STUDIES ON THE RELIABILITY OF TESTS

BULLETIN No. 12

OF THE

DEPARTMENT OF EDUCATIONAL RESEARCH

BY

ROBERT W. B. JACKSON, Ph.D. (London)

AND

GEORGE A. FERGUSON, Ph.D. (Edinburgh)

The preparation of this Bulletin was aided by a grant from the Canadian Council for Educational Research

PRICE $1.00

DEPARTMENT OF EDUCATIONAL RESEARCH
UNIVERSITY OF TORONTO
371 BLOOR STREET WEST
TORONTO 5

STUDIES ON THE RELIABILITY OF TESTS

BULLETIN No. 12

OF THE

DEPARTMENT OF EDUCATIONAL RESEARCH

BY

ROBERT W. B. JACKSON, Ph.D. (London)

AND

GEORGE A. FERGUSON, Ph.D. (Edinburgh)

The preparation of this Bulletin was aided by a grant from the Canadian Council for Educational Research

PRICE $1.00

DEPARTMENT OF EDUCATIONAL RESEARCH
UNIVERSITY OF TORONTO
371 BLOOR STREET WEST
TORONTO 5

PRINTED BY
THE UNIVERSITY OF TORONTO PRESS
1941

REPRODUCED BY
HILL LITHOGRAPHING, TORONTO
1948

FOREWORD

In my opinion, this Bulletin by Dr. Jackson and Dr. Ferguson gives the most penetrating analysis of the problem of reliability yet made. Reliability, they prove conclusively, is not the simple concept it was once considered. There is no such thing as the reliability of a test, but only the reliability of a test in a specified situation. The rather technical experimental analyses of the earlier chapters lead to the practical recommendations of Chapter VII.

In these recommendations regarding the reporting of data relating to the reliability of tests, attention is drawn to the necessity, (*a*) of estimating both the absolute and the relative accuracy of the test; (*b*) of drawing a distinction between reliability as usually understood and the internal consistency of the test; (*c*) of choosing the best method of analysing the experimental data of the test (the authors recommend the analysis of variance and covariance method in preference to a correlation technique); and (*d*) a combinatorial reliability analysis for tests made up of a battery of sub-tests.

We must confess that in our practices in the Department of Educational Research we have fallen short of these high ideals set by the authors. We shall not wear a hair shirt or sprinkle the head with ashes for we lived up to the lights we had. Now that further illumination has been given, we shall try to live up to these higher and brighter standards.

PETER SANDIFORD

University of Toronto
October 1941.

ACKNOWLEDGEMENTS

We wish to acknowledge our indebtedness to the Canadian Council for Educational Research for their generous grant-in-aid for this study. Our thanks are due also to the other members of the staff of the Department of Educational Research, in particular to Professor P. Sandiford, Professor J. A. Long, Miss K. M. Hobday and Miss M. Graham, for the assistance they have given in the preparation of this Bulletin.

ROBERT W. B. JACKSON,
GEORGE A. FERGUSON

October 1941.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

REVIEW OF LITERATURE ON TEST RELIABILITY

The term reliability was first introduced into mental test theory by Spearman in 1904. Since that time the literature dealing with test reliability has grown to comprehensive dimensions. The initial impetus given to the subject by the papers of Spearman [105, 106], Brown [7], and Kelley [57] was followed by a large number of empirical enquiries which contributed little or nothing either in the way of clarification or development to the fundamental reliability concepts. Recent publications, however, have reflected a tendency towards more rigorous examination of the fundamental assumptions involved in reliability theory, and some significant progress has been made in the application of analysis of variance methods to reliability problems.

No attempt has been made in the preparation of the present brief review of literature to include all articles that relate in one way or another to the subject-matter of test reliability. Such a task would be not only laborious but unprofitable. We have, however, included all studies which in our opinion have made significant contributions to the subject.

The Spearman-Brown Prophecy Formula

The formula commonly used for estimating increase in reliability with increase in the length of test was derived independently by Spearman and Brown, and published by them simultaneously in the *British Journal of Psychology*, October, 1910. Since that time a large number of empirical enquiries have been carried out to determine the applicability of this formula. The earliest of these studies was that reported by Holzinger [45]. He administered forms A and B of the Terman Group Test of Mental Ability, a test consisting of ten component parts, to 135 pupils, and calculated a reliability coefficient for each component by correlating the parallel components of the two forms. The average reliability of the ten components was used in the Spearman-Brown formula to determine a series of theoretical values which were compared with the obtained reliabilities of the cumulated components. These data seemed to indicate that the Spearman-Brown formula tended to over-predict the reliability of a test. This

conclusion in Holzinger's investigation is in part invalidated due to the lack of equivalence of the component parts of the test used. Consequently the assumptions on which the formula is based are obviously not satisfied.

Holzinger and Clayton [46] on the basis of data obtained from two forms of the Otis Self-Administering Test of Mental Ability, divided into one and one-half minute time intervals, and on the basis of additional spelling test data, concluded that when the test material was accurately calibrated, the Spearman-Brown formula furnished excellent prediction.

Kelley [60], using data published by Gordon [41] on the reliability of judgements of lifted weights, concluded that in this context the Spearman-Brown formula predicted with a fair measure of accuracy.

Ruch, Ackerson, and Jackson [90] conducted an empirical study of the Spearman-Brown formula using spelling test material. They concluded that when homogeneous test material is used, yielding equal standard deviation and equal reliability of component tests, the Spearman-Brown formula gave meaningful prediction.

Remmers, Shock, and Kelley [82] studied the application of the Spearman-Brown formula in predicting the reliability of any given number of judgements. They concluded that this formula predicted to within two probable errors the reliability obtained by experiment up to and including thirteen judges, the limit of their data. Remmers [83] reported that there was some evidence, although this evidence was not conclusive, to indicate that in the majority of situations in which subjective judgements were used, the Spearman-Brown formula indicated the number of judgements required for a given reliability.

Jordan [56] concluded that reliability coefficients secured by correlating odd and even items were higher than those secured by correlating scores on duplicate forms. He argued that the reliability coefficients derived by correlating the odd and even items were probably better measures of the reliability of the test because pupil variability was eliminated.

Remmers and Whistler [84] reported findings similar to those of Jordan, and discussed the influence of using reliability coefficients calculated by different methods in formulae that involve a measure of test reliability such as the formula for correcting a correlation coefficient for attenuation.

Ferguson [30] presented data which indicated that the correlations between the split-halves of tests given on the same day were higher than the split-halves of tests given on different days, and carried out a

factorial analysis in which the differences between the correlation coefficients were isolated as factors.

Wherry [124] contended that the Spearman-Brown formula, when used to estimate the reliability of a test, yielded results that contained constant and chance errors, and derived a correction formula. Although certain evidence was presented to support the use of this correction formula, the logical presumptions of its derivation are uncertain.

Other writers on the Spearman-Brown formula are Lanier [64], Slocombe [99, 100], and Thurstone [115].

The experiments carried out on the Spearman-Brown formula may be classified into two groups: (1) those that are concerned with predicting the reliability of a test lengthened any number of times, and in which no assumptions regarding the splitting of a test are made, and (2) those that are concerned with the use of the formula in estimating the reliability of a test from the correlation between the split-halves. The experiments of the first group indicate that if the assumptions are satisfied, the formula will yield accurate prediction, which of course it must. These assumptions are, firstly, that the standard deviations of component tests are equal, and, secondly, that all the intercorrelations between component tests are equal. The majority of investigations have concerned themselves with reliability coefficients estimated by this formula, rather than with the problem of determining whether the conditions for its valid use were satisfied.

In using the Spearman-Brown formula to determine the reliability of a test by boosting the correlation between the split-halves, the empirical evidence supports the conclusion that such coefficients are usually, although not always, higher than reliability coefficients obtained by correlating equivalent forms. This is due to no intrinsic fault in the Spearman-Brown formula, but rather to the process of splitting the test. The so-called split-half reliability coefficients are measures of the internal consistency of tests, and such coefficients do not always bear a one to one relationship to reliability coefficients obtained by administering equivalent forms.

The Standard Error of the Spearman-Brown Formula

A formula for the standard error of the Spearman-Brown formula was first published by Shen [94]. The publication of this paper was followed by a discussion on the correctness of Shen's formula, and a number of alternative formulae were developed. Papers were contributed to this discussion by Holzinger and Clayton [46], Shen [95], Douglass [23, 24], and Holzinger [47].

Shen's standard error formula requires in its derivation the formula

$$\sigma_r = \frac{1-\rho^2}{\sqrt{N}}.$$

Since the distribution of r for high values of ρ is very decidedly skewed, the above formula cannot be used in tests of significance or in estimation in the usual way for a correlation coefficient of high magnitude without serious error. Since the great majority of reliability coefficients, which involve the Spearman-Brown formula in their computation, are of the order .8 and .9, it is clear that Shen's formula will yield very inaccurate estimates of the required confidence intervals. It is questionable whether the literature on the standard error of the Spearman-Brown prophecy formula is of little more than historical interest.

The Kuder-Richardson Formulae

Kuder and Richardson [63] contributed an interesting development to reliability theory by deriving formulae for the estimation of reliability coefficients from statistics commonly computed in the selection of test items. This method is described as the method of rational equivalence. In the calculation of reliability coefficients by one of their more useful formulae, referred to as formula 20, the information required is the number of items, the test variance, and the sum of the item variances. On the assumption that all the items are of the same difficulty, a further formula is obtained which requires only the mean, the standard deviation, and the number of items to estimate a coefficient of reliability.

The Kuder-Richardson method of estimating the reliability of tests yields measures of internal consistency rather than measures of reliability, if reliability is understood in the test-retest sense. Coefficients estimated by the Kuder-Richardson formula 20 are superior to coefficients obtained by the split-half method, because any error due to a bias in splitting a test is eliminated. Ferguson [31] and Dressel [25] furnished different derivations of the Kuder-Richardson formula 20. Casanova [10] provided a variant of this formula adapted to computational purposes.

Battery Reliability

The reliability of test batteries was studied by Douglass and Cozens [22]. They pointed out that in computing the reliability of test batteries Spearman's formula for the correlation of sums should be used, unless the sub-tests are quite similar in measuring capacity; that

is, unless all sub-tests are regarded as parallel forms. Variations of the Spearman formula for different weighting systems are given.

Thomson [111] developed the theory of maximum battery reliability. He employed Hotelling's [49] solution to the general problem of obtaining weights that would maximize the correlation between two sets of variates to the specific problem of obtaining weights that would maximize the reliability of a test battery. He discussed also the relationship between maximum prediction of a criterion and maximum reliability.

Analysis of Variance

Jackson [52] applied analysis of variance methods and the methods of testing statistical hypotheses developed by Neyman and Pearson to the problem of determining the reliability of tests. In this paper methods are developed for treating four different problems: (1) the determination of the existence of a significant practice effect, (2) the determination of whether or not the test measures the capacity of the individuals tested, (3) the estimation of practice effect if it is found to exist, and (4) the estimation of the relative importance of the random errors of measurement with respect to the true measurement of the capacity of the individual. A new statistic termed the sensitivity of a test is introduced, denoted by the letter γ, and defined as the ratio of the standard deviation of true scores to the standard deviation of the distribution of errors of measurement. This sensitivity ratio is more informative than the usual reliability coefficient as a statistic descriptive of test efficiency, firstly because it is easier to interpret, and secondly because as the ratio of two standard deviations it exists on a scale in which the units are equal.

Jackson [53] applied the analysis of variance and covariance to problems of determining the effect of combining data from different classes on the estimates of reliability, and the conditions which must be satisfied before such results may be combined.

A very significant contribution to reliability theory was made by Hoyt [51] who developed a formula for estimating the reliability of a test by analysis of variance methods. Hoyt showed that for any particular test by subtracting the sum of squares among individuals and among items from the total sum of squares a residual sum of squares is obtained which may be used to estimate the discrepancy between the obtained variance and the true variance. These data may be used to estimate the reliability of a test. It is probable that Hoyt's work will permit of further interesting development.

Factors Influencing the Reliability of Tests

The reliability of a test is determined by a wide variety of causes. Symonds [103] listed 25 factors which influence the reliability of tests. Of these factors a number have been subject to specific study including the following: (1) the difficulty of the test items, (2) number of responses in items of the multiple-response type, (3) practice effect, (4) function fluctuation, (5) the variability of the group tested.

Influence of Item Difficulty on Test Reliability

The influence of the difficulty of test items on test reliability has been studied by Symonds [104] and Thurstone [117]. Symonds presented convincing argument to show that a test made up of items of .5 difficulty value measured an individual most accurately, and that the best test for measuring a school grade was made up of items that could be answered with 50 per cent accuracy by the average individual. Thurstone reported an investigation on the relationship between the diagnostic value of a test and the difficulty values of the items composing it, the diagnostic value being defined as the correlation between scores on sub-tests containing items at different levels of difficulty with total scores summed over all sub-tests. The conclusion was that the diagnostic value of a test, and, therefore, its reliability, was a maximum when the items were about 50 per cent difficulty. The diagnostic value was found to decrease when the difficulty of the items departed from this 50 per cent level.

Influence of Number of Item Responses on Test Reliability

Workers in the field of test construction have long realized that increasing the number of alternative responses in items of the multiple-response type increased the reliability of the test, the argument being that the reliability was increased by decreasing the probability of making a score by guessing alone.

Asker [4] approached this problem from the point of view of elementary probability. Sims and Knox [97] studied the reliability of multiple-response tests when presented orally, and found with their data that a test composed of four-response items presented orally was more reliable than a test composed either of three or five responses presented orally or five responses presented visually.

Remmers and others [84, 85, 19] formulated the hypothesis that increase in the reliability of tests with increase in the number of responses was a function of the Spearman-Brown prophecy formula, the argument being that doubling, for example, the number of re-

sponses was equivalent to doubling the length of the test. This hypothesis was tested on data of different types. Some results were inconclusive; others seemed to corroborate the hypothesis. The present writers feel that no intrinsic one to one relationship exists between the lengthening of a test and increasing the number of item responses, although such an hypothesis may in practice frequently describe the data.

Ferguson [31] derived formulae based on certain probability considerations, whereby the increase in reliability with increase in the number of alternative responses could be estimated. The assumptions upon which these formulae are based may frequently not be satisfied in practice. These formulae represent a rationalization of the problem.

Influence of Practice Effect on Test Reliability

Reliability coefficients calculated after different amounts of practice have been published by Gates [39], Gundlach [43], Slocombe [101], and Anastasi [3]. Anastasi pointed out that practice increased the effectual length of the test; consequently the item difficulty values change. Since the reliability of a test is a function of the difficulty values of the items, the reliability being a maximum when all items are of .5 difficulty, any factor which influences the difficulty values will influence the reliability.

Test Reliability and Function Fluctuation

The variability of cognitive function, sometimes termed function fluctuation, is a factor which may influence the magnitude of reliability coefficients calculated by administering equivalent forms of a test with a time interval between the two administrations. Thouless [114] derived an index for the measurement of function fluctuation. Paulsen [79] suggested that functional variability was responsible for the discrepancy between reliability coefficients calculated by the split-half method and coefficients obtained by correlating equivalent forms with a time interval between the two administrations. He suggested that the correlation between equivalent forms could be corrected for attenuation, using the split-half reliabilities in the denominator of the attenuation formula, and the coefficient thus corrected used as a coefficient of trait variability.

With verbal test material function fluctuation in the experience of the present writers seems to have very little influence on the magnitude of the reliability coefficient calculated by correlating equivalent forms of a test after a time interval. With other types of material its

influence on the reliability coefficient may be more pronounced. Furthermore the discrepancy between split-half reliability coefficients and coefficients calculated by correlating equivalent forms is attributable largely to factors other than function fluctuation.

Influence of Variability of Group Tested on Reliability Coefficient

The magnitude of a reliability coefficient is a function of the heterogeneity of the group tested. Kelley [57] developed a formula for determining the reliability of a test in one range of ability given its reliability in another range of ability. Cureton and Dunlap [14] published a nomograph to facilitate the use of this formula. Rulon [91] published a graph to serve the same purpose. Dunlap and Cureton [27] developed a formula for determining the standard error of a reliability coefficient estimated from a coefficient for a different range of ability.

Further Research on Test Reliability

Despite the fact that investigations into the reliability of tests, and closely associated topics, have been comprehensive in character, there is every indication that numerous significant developments remain to be made in this field.

Firstly, more rationalization of problems which have hitherto been approached purely by empirical methods is required. An example of this type of development may be cited. The reliability of a test is a function of the variance of scores relative to a defined population. This variance is in turn a function of the error variance, the variance of difficulty values, and the mean test score. Thus if the variance of difficulty values is decreased, the variance of scores in the defined population is increased; consequently the reliability of the test is increased. Test makers have for some time been aware of the existence of a functional relationship between the reliability of a test, the variance of test scores, the variance of difficulty values, and the mean test score, but the precise nature of this apparently complex relationship was unknown. Very recently, however, equations were derived in our Research Department which showed the precise nature of this relationship; hence we are now able to estimate the changes that will occur in the variance of scores and in the reliability when by the removal or addition of certain items the mean score is changed, or the variance of difficulty values is changed, or both are changed simultaneously. While the solution of such small problems as this may in themselves be of no great importance, such developments

are essential in the attainment of a suitable rationale underlying the methodology of test construction.

The work of Kuder and Richardson [63], and of Hoyt [51] has given rise to a number of problems which require investigation. The coefficients derived by the methods suggested by these writers are indices descriptive of the internal consistency of tests. The relationship between such coefficients and the reliability of tests obtained by test-retest methods requires to be established. In experimenting on this problem it is not sufficient to design experiments merely to show the relationship between the two types of coefficients. Care must be taken to determine the causes of any discrepancy.

Some research may profitably be carried out on the relationship between various methods of item selection and test reliability, which, while not too laborious arithmetically, will select the most reliable battery of items.

The reliability of different types of test material requires further investigation. The determination of an error variance which is in large measure a characteristic of the type of test material, and independent either of the variability of the group or of the differences in difficulty between items, should now be possible by analysis of variance methods.

The influence of functional variability on reliability coefficients determined by test-retest methods has not yet received adequate attention. Such variability may be a characteristic of the type of test material used.

The above short summary includes only a few of the possible aspects of test reliability which remain to be investigated. Some of these problems are at present being studied in the Department of Educational Research, University of Toronto, and the results will be published in the near future.

CHAPTER II

THE CONCEPT OF RELIABILITY

Measurement is as important in education as in the other sciences. In many, although not all, fields of scientific enquiry experiments are designed to demonstrate the truth or falsity of some hypothesis. Measurement of the character or characters of the population specified by the hypothesis is an essential part of the experiment. An analysis of the experimental results will, if the experiment has been designed properly, generally enable us to state whether or not the hypothesis is true—at least in the particular situation we have chosen to consider. Our statement concerning the truth or falsity of the hypothesis is based on a comparison of the experimental results with the results to be expected if the hypothesis is true. It will be seen that, even if we assume the experiment has been designed correctly and adequately controlled, the validity of our statement concerning the truth or falsity of the hypothesis will depend on the accuracy of the measurements.

We are, therefore, largely at the mercy of the measuring instrument we choose to employ. Tests or examinations are our measuring instruments in the field of education. There are many kinds of tests and examinations but they have one feature in common: namely, that they are designed to measure some ability or capacity of individuals or groups of individuals. Since, in this study, we are interested in tests only as measuring instruments, it is immaterial for our purpose whether the test is, for example, an essay examination, a new-type achievement test or an intelligence test. We are concerned with their general, not their specific, purpose, and shall speak of "tests" in the general sense of the term. The results given here are, of course, applicable to all types of tests.

If we knew the true value of a character there would be no need for measurement. It follows that when we admit the necessity of measuring some character we admit that we do not know the true value. It is not always made clear that the measurement obtained is not necessarily exactly equal to the true value. The value we obtain is only an estimate, and as such is subject to error. A measuring instrument is not perfect, and, when we use it in measuring, we make errors. Let us assume, for example, that we are measuring the length of a

piece of wood with a yard-stick. If we make several independent measurements of the object, the values obtained may not differ very much but they will not all be the same. If we repeat the measurements, using another yard-stick, we shall probably obtain very similar results. It may happen, however, that our second set of measurements are on the average larger than the first. Perhaps our second yard-stick has been graduated incorrectly, or has shrunk. In this case we should say that our second set of measurements is biased, i.e. there is a constant error in addition to the usual random errors of measurement. This bias effect is not usually classed as an error of measurement, but, from the point of view of estimating true values, it is an error effect. These two effects, bias and random errors, are generally independent of each other and will be treated as such in the following discussion.

A particular measuring instrument is not necessarily equally appropriate for use in all situations. If we were measuring a group of objects which differed in length by as much as six inches, for example, an ordinary yard-stick would probably be considered a satisfactory instrument. It would not be satisfactory, however, if our objects differed in length by not more than one-tenth of an inch. The position here is different from the one considered previously in that we are using the measurements as estimates of the true lengths of the objects and also to distinguish between the objects. An instrument which is satisfactory for one purpose may be of little value for another, or inappropriate in a different situation. We can, therefore, judge whether or not a measuring instrument is satisfactory only when we know the purpose for which it is to be used, and the conditions under which it is to be used.

The tests used as measuring instruments in education and psychology are generally more inaccurate than the instruments used in other sciences. In planning experiments and in interpreting experimental results in this field, therefore, a knowledge of the accuracy or inaccuracy with which our instruments measure is essential. The ideas discussed above in relation to physical measurements seem to the writers to be fundamental and will be applied to the problems of mental measurement in the following discussion. There is not an exact analogy, of course, but some of the basic concepts are common to both fields. Additional problems enter into mental measurement: for example, the difficulty of graduating our instruments and the possibility that the objects we measure (i.e. the individuals) may change, at least with regard to our test, either between measurements or while

being measured, or may be changed by the very process of measuring. We find also that a particular test is not necessarily equally satisfactory for measuring individuals in different groups; for example, individuals of different chronological ages or in different grades. In physical measurements, on the other hand, we can always say that a particular instrument measures with a fixed degree of accuracy irrespective of the group of objects measured. The difference seems to be due partly to the fact that tests are designed to measure ability at a particular stage in the growth of the individual. This limitation is introduced in the construction of the tests. Short, easily administered group tests are demanded, and, indeed, are the only type suitable for the ordinary testing programme. They, however, sample only a limited range of ability. We may, as is done in individual tests, arrange a series of tests which will cover a number of such stages but this introduces difficulties in the construction and use of the test and we must still determine how accurately it measures at different points. It is clear that under these circumstances the tests (group or individual) will be satisfactory measuring instruments for only a fairly well-defined range of ability. Our attention will be confined mainly to group tests but it is obvious that the results are, with a slightly different interpretation, equally applicable to individual tests.

Tests, like other instruments, may be used for many purposes. Whatever the specific purpose, the scores obtained on the test are used either as

(1) estimates of the true scores of the individuals,

or

(2) estimates of the ability of individuals relative to that of the others in the group, i.e. used in estimating the relative ability of individuals and in distinguishing between individuals or groups of individuals.

These categories are not altogether independent, and we may sometimes use the same scores for both purposes, but they are convenient for the purpose of classifying the problems which arise in determining the accuracy or adequacy of tests. The following comparison of the problems of measuring height and intelligence will illustrate the convenience of the division.

Let us consider, first, the problem of measuring the height and intelligence of an individual or, more exactly, of estimating his true height and intelligence. It is sufficient for ordinary purposes if we know an individual's height to the nearest inch and whether he is of normal, above normal or sub-normal intelligence. There is, of course,

always the difficulty of deciding border-line cases, but in most instances an ordinary measuring instrument, such as a yard-stick for measuring height and a group test for measuring intelligence, would probably be considered adequate. If a particular level of accuracy is desired and specified, for example if the height is to be measured to the nearest tenth of an inch, we must choose an instrument which satisfies this requirement. It is most important in these cases to make certain that our measurements are not biased, or, if they are biased, that the amount of bias is known in order that a correction may be applied to the original measurements. Otherwise, the measurements and the magnitude of the errors of measurement will be all the information which it is necessary to give. When we say, for example, that the height of an individual is 5 feet, 10 inches and his I.Q. rating is 130 points, we do not mean that his height is *exactly* 5 feet, 10 inches or that he has an I.Q. of *exactly* 130 points. We mean that his height is *approximately* 5 feet, 10 inches and his I.Q. is *approximately* 130 points. These may be exactly equal to the true values, they may be too high or they may be too low; we simply do not know which is the case. If we know the magnitude of the errors of measurement, however, we can specify a certain range, for example 115 to 145 I.Q. points, and state that this range will cover the true value. In making this statement we know that we shall be correct in a certain fixed proportion of cases, say in 99 out of 100 cases, depending on the degree of accuracy desired. The magnitude of the errors of measurement, or rather a comparison of these errors with the degree of accuracy required, will determine whether or not a test is adequate and satisfactory in a particular situation.

The position is very different when we consider the problem of distinguishing between individuals or groups of individuals. If, for example, the individuals differ in height by not more than one-tenth of an inch, then an ordinary yard-stick graduated in inches and quarters of an inch would not be an adequate or satisfactory instrument to use. We would not be able to rank the individuals according to height with any degree of confidence. Similarly, when we are comparing two or more groups we can judge whether our measuring instrument is adequate or satisfactory only if we know, or can determine, the size of the differences between the groups. If these are large, then even an inaccurate instrument might be satisfactory for the purpose of distinguishing between the groups although it would be of little value for the purpose of distinguishing between individuals in the same group. A test need not be very accurate, for example, if all we require

is that it will enable us to distinguish between a group of morons and a group of individuals of very high intelligence. If, on the other hand, our groups are more nearly equal in ability, our test may not be accurate enough to enable us to detect the difference between them. In this case the test would not be satisfactory for the purpose of distinguishing between either the groups or the individuals within the groups. If our groups are heterogeneous and the difference between them is small, it may happen that our test is accurate enough to enable us to distinguish satisfactorily between individuals within the groups but is not accurate enough to enable us to distinguish between the groups. In such a case we should probably conclude that there is no difference between the groups, but it would be more exact to say that there may be a difference but our test is not sufficiently accurate to enable us to detect it. The question of bias is not so important in problems of this kind. If we may assume, as we generally do, that this effect is constant for all individuals, then our comparison between individuals or groups of individuals will be unaffected. It is important only when we are interested in estimates of the true scores of the individuals or groups of individuals.

We are interested, therefore, in both the magnitude of the errors of measurement and the relation between the size of the errors and the size of the differences between the objects measured. In other words, we are interested in both the absolute and the relative accuracy of our measurements. There is also the question of bias, i.e. the constant correction to be applied to our measurements.

Definition of Reliability

The term "reliability" is used in psychological and educational work, and we customarily speak of the "reliability of a test" or other measuring instrument. According to Walker [120], this was introduced in 1910 by Spearman, who defined the term "reliability coefficient" as the (correlation) "coefficient between one-half and the other half of several measurements of the same thing." In a later work [108], Spearman defines reliability as follows:

> "——— reliability; this means the amount of correlation between two or more ratings of the *same* kind."

In the beginning, therefore, reliability and correlation were connected, and this connection has not yet been broken. Since the interpretation of correlation coefficients is rather difficult, this connection has not been a happy one and has tended to confuse rather than clarify the issue. The difficulty is that the correlation coefficient is a measure

of the degree of relationship between two (or more) variables, i.e. a measure of the *agreement*, whereas in determining the accuracy of our measurements we are more interested in the *disagreement* of the results, i.e. in a measure of the departure from a perfect relationship. These measures are related, of course, but the connection with correlation has meant an indirect rather than a direct approach to the problem.

This confusion can best be illustrated by a comparison of some of the definitions of reliability in common use. One of the definitions widely used is that given by Otis [78]:

> "The term reliability is used technically in connection with tests to mean the degree to which a test is consistent in measuring that which it measures."

A criticism of this definition is that it defines reliability in terms of consistency, which does not help very much as the term "consistent" or "consistency" must also be defined. The definition given by McCall [70] is similar:

> "By reliability of a test is meant the amount of agreement between results secured from two or more applications of a test to the same pupils by the same examiner."

The difficulty here is to determine what is meant by "amount of agreement"; this seems to be similar to the idea of the "degree of relationship" which underlies correlation.

In contrast to these, we find that Kelley [61] gives the following definition:

> "——— the question of reliability is that of how accurately a test measures the thing which it does measure."

Sandiford [93] gives a similar definition:

> "By reliability is meant the accuracy with which the test measures whatever it does measure. It is, therefore, synonymous with accuracy in measurement."

The definition given by Monroe [71] relates reliability directly to the errors of measurement:

> "The reliability of a test refers to the magnitude of the differences between the *obtained scores* and the *true scores*. These differences are the variable *errors of measurement*."

Thurstone [116] gives a similar definition:

> "A test that is subject to relatively small chance factors in its score is said to be reliable while a test with considerable variation from one occasion to another is said to be unreliable."

Here again the difficulty is to determine what is meant by "relatively small" and "considerable". Later, on page 3, Thurstone [116] distin-

guishes between the "errors of measurement" and the "relative stability of the scores", and says that the reliability coefficient describes the relative stability of the scores.

It seems clear that there is some disagreement among the various authorities on the definition of reliability. On the one hand, we find the emphasis placed on the accuracy of the test or the errors of measurement, and, on the other hand, the emphasis is placed on the consistency of the measurements or the relative stability of the scores. The necessity of considering the relation between the errors of measurement and the size of the differences between the individuals tested has been stressed by Franzen and Derryberry [36] and Jackson [52]. The latter author introduces the term "sensitivity" of a test, and suggests a different statistical method to be used in analysing the experimental results. (This problem will be considered in detail later in this Bulletin.) The comparison of the errors of measurement with the differences between the individuals tested seems to be a new approach, but it may be that this is what the other authors had in mind when speaking of consistency and relative stability; it is assumed in the following discussion that this is the case.

It is clear that we have two problems here, not just one. We must determine (1) the magnitude of the errors of measurement and (2) the relation between the size of these errors and the size of the differences between the individuals tested. It is suggested that one term, such as reliability, is not sufficient and that the position would be clarified if we stopped using such a "blanket" term and used instead the terms (defined earlier) "absolute" and "relative accuracy" of measurement. Another reason for suggesting a change in terminology is that the terms reliability and reliable are used (at least on this continent) in another and quite different sense. Garrett [38], for example, speaks of the "reliability" of the mean, meaning the errors of estimation of the mean, and also of the "reliability", i.e. the significance, of the difference between two means. These uses of the term introduce the concept of errors of sampling; it is not surprising, therefore, that the meaning of the term reliability is not clear.

There is another related concept which may be discussed briefly at this point. This is the concept designated by the term "index of reliability". According to Walker [120], this was introduced by Spearman and Abelson, although Kelley obtained independently the same result. The index of reliability (actually the square root of the reliability coefficient) is an estimate of the correlation between the obtained scores and the true scores on a test. Theoretically, this is a

useful concept and estimate, but, since we can never know the true scores, it is of little practical value. For this reason, therefore, no further discussion will be given of this concept or estimate. It may be mentioned that occasionally authors of tests give the estimates of the index of reliability when discussing the reliability of their tests. As the index of reliability is greater, and sometimes considerably greater, than the reliability coefficient, this practice is to be condemned as the values quoted give the impression (to those who are unfamiliar with this type of work) that the test is a more accurate measuring instrument than it actually is.

The position with regard to the problem of reliability may be summarized as follows: The tests and examinations used in psychology and education may be considered as measuring instruments. Any particular test or examination, like any other measuring instrument, is designed for use in particular situations and under certain well-defined conditions. If it is used in other situations or under other conditions, the measurements may be of little or no value. Even in the most favourable circumstances, however, the test or examination is not a perfect measuring instrument and in using it we make errors. It is, of course, essential for us to know the size of these errors, i.e. the accuracy or inaccuracy with which we can measure. The problem, however, is complicated by the fact that the measurements, i.e. the scores, may be used as estimates of the true scores of the individuals and, also, in distinguishing between individuals or groups of individuals. It is, therefore, necessary for us to determine the magnitude of these errors of measurement and also to obtain a measure of the size of the errors in comparison with the size of the differences in ability of the individuals between whom we wish to distinguish. It is suggested, finally, that in order to distinguish between these two aspects of the problem, we should use the terms "absolute" and "relative" accuracy of measurements instead of the single blanket term "reliability".

CHAPTER III

THE MEASUREMENT OF RELIABILITY

Before discussing the methods used, or suggested for use, in measuring the reliability of tests, it is necessary to point out that the values obtained are in all cases simply estimates of the true reliability of the test.* This is not always clearly recognized; we find, in fact, that many workers seem to regard the estimated values as the true ones and in very few cases is any attempt made to show the magnitude of the errors of estimation. As it is felt by the authors that workers will be amply repaid for the time spent in clarifying the position on this point, the next few paragraphs will be devoted to a discussion of what we mean when we speak of the "true" or "population" values, and of the importance of considering the errors of estimation.

When we speak of the true or population value of a statistic, such as a reliability coefficient, we have in mind some hypothetical group or population of cases. Generally, the number of cases in the population is considered to be infinite, or at least very large. The true or population value of the statistic is a parameter (it may be the only one, or one of many such parameters) determining the distribution of the cases in this hypothetical group. When we speak of the errors of measurement involved in our measurement of the length of a steel rod, for example, we think of the differences between the true length of the rod and the values we obtain when we measure the length with some measuring instrument. We can, of course, make many measurements and, if we knew the true length of the rod, we would have a distribution of such differences. The population of differences would be formed of the differences obtained from an infinite number of such measurements; these differences, subject to certain assumptions, would be normally distributed about zero with a constant standard deviation, say σ. The distribution of the differences in the population is, in this case, determined solely by the value of this parameter, the constant standard deviation σ. It follows that if we knew the value of σ, we would know the magnitude of the errors of measurement. In practice, of course, we cannot make an infinite number of measurements so the true value of σ is unknown; what we do is to estimate σ. We make a certain

*See also Appendix A.

number of measurements and calculate the standard deviation, say S, of the distribution of these and use this as our estimate of σ. It is clear that S may not be exactly equal to σ; we never do know exactly the magnitude of our errors of measurement.

The small group of measurements which we made, form what is termed a sample, i.e. a certain number of all the possible measurements which could be made. In experimental work we almost invariably work with samples and sample values, not with populations and population values. As mentioned above, however, we sometimes forget this and speak of, and use, the sample values of the statistics as though they were the true values of the parameters in the population from which we are sampling. These sample values are the only values we have, of course, so we must use them, but we should always remember that they are only estimates of the true values.

The theory of sampling enables us to take one further step. Using the estimates calculated from the sample, we may determine an interval and make the statement that this interval covers the true value of the parameter. The property of this confidence interval, as it is called, is that we know we shall be correct in making this statement in a certain fixed proportion of such cases [77]. This is, however, as far as we can go; we simply do not know, and in most cases cannot hope to know, the true value of the parameter in the population from which we are sampling.

This idea of estimating the population value from the value calculated for a sample has an important bearing on the whole problem of reliability. In quoting the sample value, we claim, explicitly or implicitly, that it is an estimate of the parameter in the population from which we are sampling. It follows that we have in mind some particular population and, to be consistent, we should specify the population to which our estimate refers. Let us assume, for example, that an author of a test states that the value of the reliability coefficient for his test is 0.9. What are the possible populations to which this sample value may refer? This depends, of course, on the group or groups from which the sample was drawn. It may be that all the children tested were from one grade, or were in one class (possibly selected) of pupils in a particular grade, or perhaps the sample included all the children in one or more grades in a particular school. We must remember also that possibly the school contains pupils drawn only from a particular social or racial class, or possibly from all classes, and we may have a narrow or broad range of ability. In other cases, of course, the author may have confined his attention to children of a

particular chronological age-group, ten-year-olds, for example. It will be realized that the number of possible groups is very large, and unless the author gives us some idea of what group or groups were sampled, the value he quotes will, at least to a certain extent, be meaningless.

Further discussion of this point must be postponed until we come to consider the experimental and statistical methods used in measuring, and the kind of information which it is suggested should be given when reporting the reliability or accuracy of a test. It may be noted, however, that the population we consider is determined partly by the nature of the test, the way in which it is used, and the kind of situations in which it is used. In an individual test, for example, it is likely that a knowledge of how accurately the test measures at a particular chronological age-level would be most useful; the interest here is centred in the determination of the ability of the individual, not so much in distinguishing between individuals, although this is also of interest. In a group test, on the other hand, which is generally given by a teacher to all the pupils in a particular class or grade, information regarding the accuracy with which the test measures and also how well it distinguishes between the individuals within the group will be necessary. The statistical method suggested for use in problems of this kind gives us the answers to all these questions in a single analysis. It does not, of course, tell us which units we should use, but it does tell us what happens when we use different units or samples from different populations.

(1) *Experimental Methods*

Most authors report some statistics regarding the accuracy with which their test measures, but there are still some who ask us to accept their offered measuring instrument on faith. It seems necessary, therefore, to point out that the only way to determine how accurately the test measures is to try it out—and this is the task of the author of the test.

In determining the accuracy of a physical measuring instrument, the same objects are measured several times and from the differences found we may calculate an estimate of the errors of measurement. Possibly the easiest and simplest method to use in determining the accuracy of a mental measuring instrument is the same, i.e. to repeat the test on the same group of individuals after a certain period of time has elapsed. In this case, also, we may obtain an estimate of the errors of measurement from the differences found between the two sets of measurements. The problem, however, is not as simple in the case

of mental as in physical measurements. In the latter case the objects, except in the few cases where their destruction is involved, are unaffected by the measuring process and remain unchanged over a short or long period of time. In mental measurements, unfortunately, the objects measured—the children—react to the process of measurement and in any case change over a period of time, at different rates of change. We cannot do much about this, however, since the children are alive. It is simply an additional obstacle peculiar to the field of measurement whenever living organisms are involved. It is one of the reasons why this simple test-retest method, as it is called, is not used in all experiments concerned with the determination of the reliability of tests.

The second experimental method requires two or more equivalent forms of the same test. Instead of making just one test, two or more forms of it are constructed; these are matched, generally item by item, for difficulty, content, etc. For tests for which alternative forms are available, we do not repeat the same test but give one of the alternative forms on the second trial. If the forms are truly equivalent, this method is to be preferred as it overcomes some of the weaknesses of the test-retest method. As the alternative forms may not be exactly equivalent, however, this procedure may, in certain cases, introduce an additional disturbing factor. Since, however, the time elapsing between the giving of the two forms of tests may be very short, the method may be said to obtain children whose abilities are more or less constant on the two trials. In using this method, we again obtain an estimate of the errors of measurement from the differences found between the two sets of measurement. It will be seen that we assume that these differences, except for a constant practice effect, are caused entirely by the errors of measurement. It follows that if the two forms are not equivalent, these differences will tend to be increased and to this extent our estimate of the errors of measurement will be biased. We would, in fact, conclude that our test is more inaccurate than it actually is. It will be agreed, however, that, if err we must, it is better to overestimate than underestimate the magnitude of the errors of measurement.

The third method consists simply in giving the test once and from the scores obtained, or some function of them, estimating how accurately the test measures. This method seems to overcome all the weaknesses mentioned above, and from a theoretical point of view, is probably to be preferred. It will be shown later, however, that in practice the assumptions underlying the statistical methods used in

analysing the results obtained by the use of this method are not always satisfied. The validity of our estimate of the accuracy of our test depends mainly upon the type of test we have and, to a lesser degree, the relative difficulty and position of the individual items. From the practical point of view, therefore, this method is probably the weakest of all, and it is suggested that it should be used only for cases in which neither of the other methods may be employed.

(2) *Statistical Methods*

Most of the statistical methods used in estimating the reliability of tests, or the accuracy with which they measure, may be applied without change to the results obtained by using any of the experimental methods. There are, however, certain methods which are developed for use in particular situations. The analysis of variance method suggested by Jackson [52], for example, cannot always be used in analysing the results obtained from the third experimental method. The author suggested that it could be used in all cases, subject to certain possible differences in interpretation, but it has since been found that some of the assumptions underlying his method are not always satisfied. This point will be discussed in more detail later; it is sufficient at this stage to point out that it may be used in some but not all of these cases.

It is convenient to divide the discussion on these statistical methods into three parts and to deal with each one separately:

(*a*) methods used in estimating reliability coefficients;
(*b*) methods used in estimating the errors of measurement;
(*c*) the method suggested for use by Jackson (see above) in estimating the sensitivity of a test.

(a) *Methods used in estimating reliability coefficients*

Although it gives us only an indirect measure of the accuracy of the test, the reliability coefficient is the most widely used measure of reliability. It is a correlation coefficient, the coefficient for the two sets of test scores obtained from the same group of individuals. If the test is given only once, it is customary to take the scores on the odd and even items and use the obtained correlation coefficient as an estimate of the reliability of either half of the test. The estimate of the reliability for the whole test is then calculated by using the well-known Spearman-Brown formula for double length. If we denote by X_i the score obtained by the i-th individual on the first testing; by Y_i the score obtained by the same individual on the second testing;

by Σ summation over all N values of i, then the reliability coefficient, r, may be defined as

$$r = \frac{\Sigma X_i Y_i - \frac{(\Sigma X_i)(\Sigma Y_i)}{N}}{\sqrt{\left\{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N}\right\}\left\{\Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{N}\right\}}} \quad \ldots\ldots\ldots (1)$$

If the coefficient so obtained refers to only half the test, then the coefficient for the whole test is obtained by using the formula

$$r_w = \frac{2r_{\frac{1}{2}}}{1+r_{\frac{1}{2}}} \quad \ldots\ldots\ldots (2)$$

where $r_{\frac{1}{2}}$ denotes the half-test, and r_w the whole test, coefficient, respectively. It may be noted, however, that under certain conditions the value obtained by using equation (1) is not the best estimate of the population value; this particular problem is discussed in detail in Appendix A of this bulletin. This appears to be another of the cases in which research workers have accepted a statistical method without examining critically the assumptions underlying it and comparing these with the conditions of their own problems.

Kuder and Richardson [63] have suggested another method of estimating the reliability coefficient. They have shown that, subject to certain assumptions, we may obtain an estimate of the reliability coefficient from the results of a single application of the test by using the following formula

$$r = \frac{n}{n-1}\left\{\frac{S^2 - \Sigma pq}{S^2}\right\} \quad \ldots\ldots\ldots (3)$$

where r is the reliability coefficient; n the number of items in the test; S the standard deviation of the distribution of scores obtained on the test; p the proportion of subjects passing any given item; $q = 1 - p$; Σpq the sum of the products of p and q for all items in the test. This is one of the special methods mentioned above and, as will be shown later, it seems to give a measure of the internal consistency rather than the reliability of the test in the usual sense of the term.

(b) *Methods used in estimating the errors of measurement*

The reliability coefficient by itself does not give us an estimate of what has been termed the absolute accuracy of our measurements, although, as will be shown in part (c) of this section, it does give us an indirect estimate of their relative accuracy. We may, however, use the reliability coefficient in obtaining our estimate of the absolute

accuracy, i.e. in calculating an estimate of what is termed the standard error of measurement of an individual score. This estimate, however, may be calculated directly from the distribution of differences between the scores obtained on the two testings. Assuming that these differences, except for the constant practice effect, are caused solely by errors of measurement, we may use $1/\sqrt{2}$ times the standard deviation of the distribution of differences as a measure of the errors, and, therefore, speak of the "standard error" of measurement. The other method is to use the standard deviation of the obtained test scores times $\sqrt{1-r}$ as the estimate, i.e

$$S_E = S\sqrt{1-r} \qquad \ldots\ldots\ldots(4)$$

where S_E denotes the standard error of measurement; S the standard deviation of the distribution of scores on the test, and r the reliability coefficient. If the standard deviations of the distributions of the two sets of scores are equal, then these two values of S_E are identical.

This assumption of equal standard deviations is fundamental to all our work on the determination of the reliability or accuracy of our test. If we are using two alternative forms of a test, for example, and find that the standard deviations are different, we cannot estimate either the reliability or the errors of measurement. We must conclude, in such a case, that the two forms of the test are not equivalent and refrain from using this particular experimental method. It may be argued that the inequality of the standard deviations will not affect our correlation coefficient. This may be true but it does not affect the issue at stake. If the forms are not equivalent, as they cannot be if the standard deviations are significantly different, then they must differ in some respect. Hence we cannot use this particular experimental method, as the basic assumption underlying this method is that the two forms are equivalent. It does not matter so much here if one of the forms is easier or more difficult than the other, provided the standard deviations are the same, as this constant difference will affect none of the results except the norms for the test.

There is one danger connected with the use of this indirect method of determining S_E. The formula given in equation (4) may be used to determine S_E if, and only if, the S and r on the right-hand side refer to the same group of cases. We generally find that S_E is constant for most groups and, if this is so, then r and S must be related and hence it may be incorrect to use in the same formula an S and an r which refer to different groups. Occasionally we find that S_E itself may change from group to group (as in groups of pupils chosen from different school grades, for example), and we conclude that the test does

not measure with equal accuracy at all levels. In such cases it is, of course, particularly important to exercise care in the use of the formula given in equation (4). Since the method discussed in the next part uses a direct rather than an indirect approach to this problem, it is suggested that it is probably the better one to use.

(c) *The method suggested for use in estimating the sensitivity of a test*

The method suggested for use is the one known as the Analysis of Variance. As correlation is not used, it will seem strange to many workers in education and psychology, but it is being more and more widely used in these fields. As the general use of this method has been discussed elsewhere [53] only its application to the problem of reliability will be considered here. The theory underlying the method will not be discussed in any detail as it is felt that research workers will be interested in the practical rather than the theoretical aspects of the problem.

The idea underlying the method is very simple. It is assumed that the score made by an individual on a test may be considered as a sum of independent components, and the analysis is designed to give a measure of the influence of each of these. In the problem of reliability the factors are few in number and the results obtained in the analysis are easily interpreted. As it is easier to work with a particular situation in mind, let us start with some experimental results. The data given in the second and third columns of Table I refer to the scores made by a small class of 29 pupils on two forms of an intelligence test. What factors or components may be important? In the first place, it is clear that not all the individuals in the class are of the same ability, so we must find out how well our test measures the differences between the pupils, i.e. distinguishes between the individuals. Secondly, it will be seen that the pupils make, on the average, higher scores on Form B than on Form A; Form A was given to the children first, so this constant difference is called a measure of the "practice" effect. Finally, it will be seen that even after allowing for the influence of this practice effect, the scores on the two forms differ considerably. These residual differences we assume to be due to the errors of measurement by means of the test used, or, rather, we define the errors of measurement in this way. There is no other factor, except possibly fluctuations in the ability of the individuals and differences between the two forms, which we cannot isolate in such a simple experiment, so we class all these residual differences as error.

The next problem is to determine how to measure the effect of

these various factors or components. This is done by obtaining a measure of the amount each of these contribute to the total variance (variance is a technical term used to denote the square of the standard deviation) of the scores. We break up the total variance, or rather the sum of squares of the deviations about the mean from which an estimate of the variance is calculated, into components which we may assign to these different factors. This gives us a means whereby we may determine the importance of the influence of each of these factors, and hence draw conclusions concerning the usefulness of our test as a measuring instrument.

As far as the arithmetic of the analysis is concerned, this is quite simple—probably even simpler than the arithmetic procedure involved in the calculation of a correlation coefficient. We calculate for each individual the sum of his scores, and the difference between his scores, on the two forms of the test as shown in the last two columns of Table I. Then we calculate the sum and sum of squares of the values in each column and write these in the spaces provided in the two bottom rows of the table. In the column headed X in the table, for example, the sum is simply the total of the 29 values in the column, and the sum of squares (in the bottom row) is the total of the squares of each of the 29 values. It will be seen that three checks on the accuracy of our work may be made at this stage: the sum of $X+Y$ must be the same as the sum of X plus the sum of Y, i.e.

$$1390 \equiv 633 + 757$$

Similarly, for the differences,

$$-124 \equiv 633 - 757$$

and, for the sums of squares,

$$78760 + 1684 \equiv 2(16537 + 23685)$$

A final check is made at a later stage in the analysis.

To calculate the sum of squares, from which the estimates of variance attributable to each factor are obtained, we proceed as follows:

(1) for Error

$$\tfrac{1}{2}\left[1684 - \frac{(-124)^2}{29}\right] = 521.667$$

(2) for Between Individuals

$$\tfrac{1}{2}\left[78760 - \frac{(1390)^2}{29}\right] = 6{,}067.931$$

(3) for Practice Effect

$$\tfrac{1}{2}\left[\frac{(-124)^2}{29}\right] = 320.333$$

TABLE I

SCORES RECEIVED BY PUPILS ON FORMS A AND B OF AN INTELLIGENCE TEST

Pupil No.	Score on Form A X	Score on Form B Y	Sum of Scores $X+Y$	Difference $X-Y$
1	9	14	23	−5
2	15	22	37	−7
3	9	12	21	−3
4	10	19	29	−9
5	40	37	77	3
6	13	8	21	5
7	19	20	39	−1
8	17	34	51	−17
9	18	19	37	−1
10	15	20	35	−5
11	24	29	53	−5
12	24	24	48	0
13	13	28	41	−15
14	29	30	59	−1
15	13	16	29	−3
16	23	26	49	−3
17	19	28	47	−9
18	24	15	39	9
19	16	16	32	0
20	41	46	87	−5
21	35	30	65	5
22	24	30	54	−6
23	33	53	86	−20
24	24	27	51	−3
25	32	41	73	−9
26	20	24	44	−4
27	45	56	101	−11
28	12	11	23	1
29	17	22	39	−5
Sum	633	757	1390	−124
Sum of Squares	16537	23685	78760	1684

(4) for Total

$$16537+23685-\frac{(1390)^2}{58} = 6{,}909.931$$

It is customary, and also convenient, to put all these values in a table of the form shown in Table II.

TABLE II

ANALYSIS OF VARIANCE OF SCORES MADE BY PUPILS ON TWO FORMS OF AN INTELLIGENCE TEST

Variance	Degrees of Freedom	Sum of Squares	Mean Square
Due to Practice Effect	1	320.333	320.333
Between Individuals	28	6,067.931	216.712
Error	28	521.667	18.631
Total	57	6,909.931	.....

As the total of the sum of squares for (1), (2) and (3) must be identically equal to (4), this gives us a final check on the accuracy of the calculations.

The first column of Table II is self-explanatory; we simply list the factors in which we are interested, and the quantities in the third column have been explained above. The entries in column 2, headed "Degrees of Freedom", will require some explanation. These quantities are used as divisors in calculating the values shown in the last column (headed "Mean Square"); in the Between Individuals row, for example, $216.712 = \frac{6,067.931}{28}$, and the other mean square values are calculated in a similar manner. Generally, in calculating the value of the square of the standard deviation, we divide the sum of squares by the number of observations in the sample. In small samples, however, this estimate is biased and the bias may be compensated by dividing by the number of degrees of freedom instead of the number of observations. In examples of the kind considered here, the number of degrees of freedom to be used are as follows:

(1) Due to Practice Effect	1
(2) Between Individuals	$n-1$
(3) Error	$n-1$
(4) Total	$2n-1$

where n denotes the number of individuals tested. It will be noticed that the additive property discussed above in connection with the sums of squares applies also to the degrees of freedom, i.e.

$$1+(n-1)+(n-1)\equiv 2n-1$$

For the benefit of those who are interested in the question of the number of degrees of freedom associated with any particular sum of squares, it may be pointed out that this number is always equal to the number of independent deviations which are used in the calculation of the associated sum of squares. For the error row, for example, we see from Table I that while there are 29 differences (one for each pupil), yet in calculating the sum of squares for error we subtract the mean difference $\left(-\frac{124}{29}\right)$ from each. We have, therefore, only 28 of these difference deviations which are independent, and hence the number of degrees of freedom is 28.

With regard to the use which is to be made of the results shown in Table II, we see that in the first place we may obtain an estimate of the standard error of measurement of an individual score, denoted by S_E in equation (4), by taking the square root of the error mean square. We find $S_E = \sqrt{18.631} = 4.32$ score units. This gives us directly an estimate of the absolute accuracy of our measurements.

The next problem in which we are interested is to determine whether or not the practice effect is significant, i.e. significantly different from zero. If the practice effect is zero, then the corresponding mean square in the table will be of the same order of magnitude as the error mean square; if the practice effect is significant, its mean square will be larger than the error mean square. It follows, therefore, that if we can show the practice effect mean square is significantly greater than the error mean square, then we may conclude that the practice effect is significant. In making this test, we proceed as follows:

(1) calculate

$$F = \frac{320.333}{18.631} = 17.19 ;$$

(2) refer to Snedecor's table [53] of F with degrees of freedom $n_1 = 1$ and $n_2 = 28$;

(3) conclude that the two mean squares are of the same order of magnitude if the calculated value of F is less than the 5% (or 1%) point of the distribution of F given in the table, or conclude that the two mean squares differ significantly if the calculated value of F is greater than the 5% (or 1%) point of the distribution of F given in the table.[1]

[1]The question of whether to use the 5% or the 1% point of the distribution of F as a critical value is a personal one, and is sometimes determined by the nature of the problem under consideration. It is customary, however, to conclude that the mean squares are different if the calculated value of F is greater than the

In our particular case we find that the 5% and 1% points of the distribution of F are 4.20 and 7.64, respectively. As the value of F ($F=17.19$) which we obtained is considerably greater than the 1% point we conclude that the practice effect mean square is greater than the error mean square, and hence that the practice effect is significant.

The next problem is to determine whether or not the test measures with an accuracy sufficient to enable us to distinguish between the individuals tested. If the accuracy with which the test measures is not sufficient for this purpose, then the differences between the scores obtained by the individuals on the test will be very small and due solely to the errors of measurement; in this case the between individuals mean square will be of the same order of magnitude as the error mean square. If, on the other hand, the accuracy with which the test measures is sufficient for this purpose, then the differences between the scores will be larger than could be explained solely on the basis of the errors of measurement, and the between individuals mean square will be significantly larger than the error mean square. It follows, therefore, that if we can show the between individuals mean square is larger than the error mean square, then we may conclude that the test measures with an accuracy sufficient to enable us to distinguish between the individuals tested. The procedure followed in making this test is similar to that considered in the previous case:

(1) calculate

$$F=\frac{216.712}{18.631}=11.63\ ;$$

(2) refer to Snedecor's table of F with degrees of freedom $n_1=n_2=28$;

(3) conclude that the two mean squares are of the same order of magnitude if the calculated value of F is less than the 5% (or 1%) point of the distribution of F given in the table, or conclude that the two mean squares differ significantly if the calculated value of F is greater than the 5% (or 1%) point of the distribution of F given in the table. As the value we found ($F=11.63$) is larger than the 1% critical value, we conclude that the two mean squares differ significantly and hence that the accuracy with which our test measures is sufficient to enable us to distinguish between the individuals tested.

1% point, to refrain from drawing any conclusion if it lies between the 5% and the 1% points, and to conclude that they are of the same order of magnitude if it is less than the 5% point. This is, however, a purely arbitrary choice of a critical value.

This leads us to the problem of determining the relative accuracy of our measurements, i.e. the relation between the size of the errors of measurement and the size of the differences between the individuals tested. Jackson [52] has suggested that the simplest measure of the relative accuracy is what he has termed the sensitivity of the test. This is defined as

$$\gamma = \frac{\sigma_c}{\sigma} \qquad \ldots\ldots\ldots (5)$$

where γ denotes the sensitivity of the test; σ_c the standard deviation of the distribution of ability in the population from which we are sampling, and σ is the standard deviation of the distribution of the errors of measurement. It will be seen that this measure is particularly easy to interpret; if γ is small, then the errors of measurement will be large in comparison with the differences between the abilities of the individuals tested, and the score obtained by an individual on the test may be determined largely by these random errors of measurement. For a particular value of γ, we can determine the probability, say η, from the tables of the normal integral of making an error greater than or equal to σ_c units due to chance alone in using the score as an estimate of the ability of an individual. Certain values of γ and η are given in Table III. It will be seen that even for fairly large values of γ, the random errors of measurement may be very important in determining the actual score of an individual on the test.

TABLE III

VALUES OF γ AND η

γ	0.5	1.0	1.5	2.0	2.5	3.0
η	.62	.32	.13	.046	.012	.003

In the population from which we are sampling, the sensitivity and the reliability coefficient are related, i.e.

$$\gamma = \sqrt{\frac{\rho}{1-\rho}} \qquad \ldots\ldots\ldots (6)$$

where ρ denotes the population reliability coefficient. The reliability coefficient does, therefore, give us an indirect estimate of the relative accuracy of measurements. In estimating γ from the sample values, however, it is better to proceed directly rather than to attempt to use the reliability coefficient. The estimate of γ discussed in the next

paragraph is the best estimate which can be obtained from the sample values.

In estimating γ, we may use either a unique estimate or, better still, the "confidence interval" which was discussed earlier in this chapter. To obtain the unique estimate, we proceed as follows:

(1) subtract the error mean square from the between individuals mean square;

(2) divide the difference by twice the error mean square;

(3) use the square-root of the quotient as an estimate of γ.

For our example, from the values given in Table II, we have

(1) $216.712 - 18.631 = 198.081$

(2) $\dfrac{198.081}{37.262} = 5.316$

(3) est. $\gamma = \sqrt{5.316} = 2.31.$

To find the confidence interval we proceed as follows:

(1) calculate the ratio of the between individuals to the error mean square, which we may denote by F ;

(2) from Snedecor's table of F find the 5%, or 1%, point of the distribution of F, which may be denoted by $F_{5\%}$ or $F_{1\%}$;

(3) to find the lower limit of the interval, say $\underline{\gamma}$, using $F_{1\%}$ for example, calculate

$$\underline{\gamma} = \sqrt{\frac{F}{2F_{1\%}} - \frac{1}{2}} \qquad \ldots\ldots\ldots (7)$$

(4) to find the upper limit of the interval, say $\overline{\gamma}$, using $F_{1\%}$ for example, calculate

$$\overline{\gamma} = \sqrt{\frac{FF_{1\%} - 1}{2}} \qquad \ldots\ldots\ldots (8)$$

(5) we may make the statement

$$\underline{\gamma} \leq \gamma \leq \overline{\gamma} \qquad \ldots\ldots\ldots (9)$$

and we know that the probability of this statement being correct is .98 (it would be .90 if we had used $F_{5\%}$). For our example, we have

(1) $F = \dfrac{216.712}{18.631} = 11.63$

(2) $F_{1\%} = 2.50$

(3) $\underline{\gamma} = \sqrt{\dfrac{11.63}{5.00} - \dfrac{1}{2}} = 1.35$

(4) $\bar{\gamma} = \sqrt{\frac{29.075-1}{2}} = 3.75$

(5) $1.35 \leq \gamma \leq 3.75.$

From a theoretical point of view, the confidence interval method is the better one to use, but in practice the unique estimate is probably more convenient. In using the latter estimate, however, one must always remember that it is subject to error.

In some problems where a particular degree of sensitivity is required, it is necessary to determine whether or not the test considered reaches this standard. This problem is discussed in detail in Appendix B; the solution given is simple and involves merely an extension of the method used earlier in this section. The statistical problem proposed and solved is that of developing a test of the hypothesis $\gamma = K$, where K is some value fixed in advance. The usefulness of this statistical test is obvious.

If we use the correlation method in analysing the above experimental results, we find a reliability coefficient of $r = 0.84$. The difficulty of interpreting this result will be admitted by all. It is suggested, therefore, that it is better to use the analysis of variance method in analysing the results of experiments of this kind.

It was mentioned earlier that our results will be different if we use different groups in our experimental work, i.e. if we sample from a different population. Let us consider, for example, the effect on the results if we sample from four grades instead of just one. Figure 1 shows the distribution of scores made by pupils on two forms of an intelligence test: as different symbols are used to represent the scores which refer to different grades, this diagram shows the effect of using the broader unit in sampling. It will be seen that the values are spread along a line, like beads on a wire. The more grades we include the more important does this "elongation" effect become. It is obvious that the reliability coefficient will be increased, and in some cases greatly increased, by this effect. If we use the analysis of variance and covariance method in analysing our results, we obtain a measure of the influence of this effect, and also a measure of the reliability freed from the influence of this effect. It is clear that in this case the factor causing the "elongation" is the differences between the grades, so this is the factor which we must measure and eliminate.

In order to make this analysis we need, for each form and for each grade, the sum, sum of squares and finally the sum of products of the scores: the data relating to this example are given in Table IV.

DISTRIBUTION OF SCORES MADE BY PUPILS ON TWO FORMS OF AN INTELLIGENCE TEST

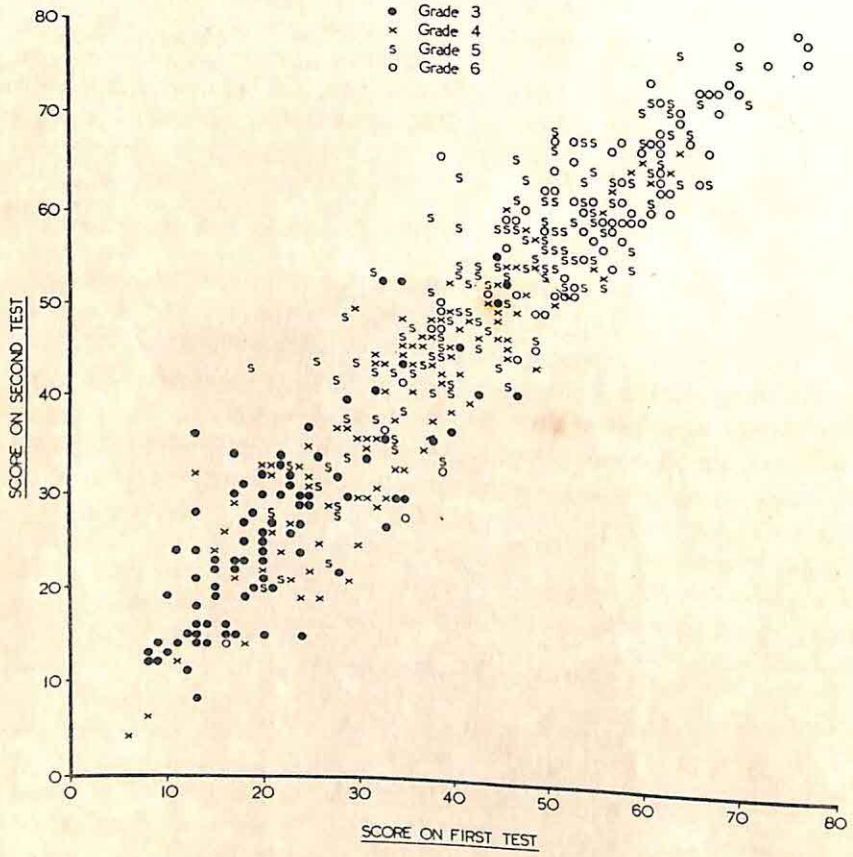

FIGURE 1

TABLE IV

DATA RELATING TO THE SCORES OF PUPILS ON TWO FORMS OF AN INTELLIGENCE TEST

Grade	Number of Pupils	Sum of Scores on First Form	Sum of Scores on Second Form	Sum of Squares of Scores on First Form	Sum of Squares of Scores on Second Form	Sum of Products of Scores on First and Second Form
3	98	2,094	2,552	53,408	77,042	62,518
4	112	3,862	4,408	150,250	193,600	168,958
5	136	6,134	7,143	295,016	394,941	338,963
6	108	5,846	6,442	329,678	397,728	360,596
Total.....	454	17,936	20,545	828,352	1,063,311	931,035

As we want the sums of squares and products of the deviations from the means, we must again calculate these separately for each grade and form, and also for the total. For Grade 3, for example, we have:

(1) *Sum of Squares of Deviations for First Form*

$$53{,}408 - \frac{(2{,}094)^2}{98}$$

$$= 53{,}408 - 44{,}743.224 = 8{,}664.776$$

(2) *Sum of Squares of Deviations for Second Form*

$$77{,}042 - \frac{(2{,}552)^2}{98}$$

$$= 77{,}042 - 66{,}456.163 = 10{,}585.837$$

(3) *Sum of Products of Deviations*

$$62{,}518 - \frac{(2{,}094)(2{,}552)}{98}$$

$$= 62{,}518 - 54{,}529.469 = 7{,}988.531.$$

It is convenient in an analysis of this kind to show all these values as in Table V. The values shown in the second row from the bottom of the table are the sums of the values in the preceding four rows. Some of these are used later in calculating a measure of the effect of the differences between grades, and the others (in the last three columns) give us the total sums of squares and products for within grades.

TABLE V

QUANTITIES REQUIRED IN THE CALCULATION OF THE BETWEEN GRADES AND WITHIN GRADES SUMS OF SQUARES AND PRODUCTS

Grade	Correction factors: Squares and Products of Sums of Scores divided by Number of Pupils			Sum of Squares and Products of Deviations about Means		
	First Form	Second Form	Products	Sum of Squares for First Form	Sum of Squares for Second Form	Sum of Products
3	44,743.224	66,456.163	54,529.469	8,664.776	10,585.837	7,988.531
4	133,170.036	173,486.286	151,997.286	17,079.964	20,113.714	16,960.714
5	276,661.441	375,165.066	322,170.309	18,354.559	19,775.934	16,792.691
6	316,441.815	384,253.370	348,703.074	13,236.185	13,474.630	11,892.926
Sum..	771,016.516	999,360.885	877,400.138	57,335.484	63,950.115	53,634.862
Total for all grades	708,590.520	929,729.130	811,663.260	119,761.480	133,581.870	119,371.740

Using these results, we may present our final analysis of variance and covariance in the form shown in Table VI.

TABLE VI

ANALYSIS OF VARIANCE AND COVARIANCE OF SCORES MADE BY PUPILS ON TWO FORMS OF AN INTELLIGENCE TEST

Variance	Degrees of Freedom	Sum of Squares First Form	Sum of Squares Second Form	Sum of Products
Between Grades.....	3	62,425.996	69,631.755	65,736.878
Within Grades.....	450	57,335.484	63,950.115	53,634.862
Total.....	453	119,761.480	133,581.870	119,371.740

The sums of the squares and products are obtained as follows: for within grades, we use the last three values given in the second row from the bottom of Table V; for total, we use the last three values given in

the last row of Table V; for between grades, using the values shown in the last two rows and columns 2, 3 and 4 of Table V, we find

(1) for the First Form
$771{,}016.516 - 708{,}590.520 = 62{,}425.996$

(2) for the Second Form
$999{,}360.885 - 929{,}729.130 = 69{,}631.755$

(3) for the Products
$877{,}400.138 - 811{,}663.260 = 65{,}736.878.$

As a check on the accuracy of our calculations, we note that in each column of Table VI the sum of the values for between and within grades is identically equal to the total given in the last row. The degrees of freedom shown in Table VI are found as follows: since there are 4 grades, the number of degrees of freedom for between grades will be $4-1=3$; the number of degrees of freedom for within grades may be obtained from the second column of Table IV: $(98-1)+(112-1)+(136-1)+(108-1)=450$; the number of degrees of freedom for the total is, from Table IV: $454-1=453$.

From the values given in Table VI, we may calculate estimates of three reliability coefficients:

(1) for between grades

$$r = \frac{65{,}736.878}{\sqrt{(62{,}425.996)\ (69{,}631.755)}} = 0.997$$

(2) for within grades

$$r = \frac{53{,}634.862}{\sqrt{(57{,}335.484)\ (63{,}950.115)}} = 0.886$$

(3) for total (i.e. all grades)

$$r = \frac{119{,}371.740}{\sqrt{(119{,}761.480)\ (133{,}581.870)}} = 0.944.$$

The difference between these last two estimates gives us a measure of the effect of using the larger unit in sampling, i.e. sampling from 4 grades instead of just one. The estimate calculated from the totals for all grades is increased by the inclusion of the very significant differences between grades, i.e. the "elongation" effect shown graphically in Figure 1. The first estimate, $r=0.997$, refers to the means for the grades and not to the individual scores; it is, as one would expect, considerably higher than the within grades estimate. When we use

the estimate for the total, $r = 0.944$, both the between grades and within grades are included; we have, therefore, neither the one nor the other but a kind of compound of the two.

This raises the question:—which estimate should be used? Unfortunately, this cannot be answered. The question as to which estimate is appropriate will be determined by the conditions of the problem on which we are working. It is probably safer to give them all, so that other research workers will have no difficulty in interpreting the results. It is clear, however, that our results may be misleading or meaningless unless we state clearly the nature of the population from which we have sampled and to which the values refer.

The above suggested analysis refers to the estimates of the reliability coefficients. We may, however, extract considerably more information from the data if we use analyses of the type shown in Table II. We first analyse the results separately for each grade, as shown in Table VII, and then compare the values in the different rows in order to determine whether or not the results may be combined [see 53, pp. 83-96]. We find that our estimates of the errors of measurement do not differ significantly from grade to grade (see the "Error" row of Table VII), so we conclude that the test measures with the same absolute accuracy at all levels. The best estimate of the standard errors of measurement, S_E, may be found from the values given in this error row of Table VII as shown below:

$$S_E = \sqrt{\frac{(1{,}636.77)+(1{,}636.13)+(2{,}272.55)+(1{,}462.48)}{97+111+135+107}}$$

$$= \sqrt{\frac{7007.93}{450}}$$

$$= 3.95 \text{ score units.}$$

When we consider the question of the relative accuracy of the measurements, i.e. compare the relative accuracy with which the test measures in different grades, we find significant differences. The test distinguishes between the individuals best in Grade 4, the efficiency is slightly lower in Grades 5 and 6, and poorest in Grade 3. The test seems to be too hard for Grade 3 and hence does not distinguish between the individuals so well. It is clear that the test is better suited for testing children in Grades 4, 5 and 6.

With regard to the practice effect, we find again that the differences between the grades are significant—mainly due to the large practice effect occurring in Grade 5. No explanation of this could be found, however, so we can only conclude that, for some unknown reason, the

TABLE VII

ANALYSIS OF VARIANCE OF SCORES OF PUPILS ON TWO FORMS OF AN INTELLIGENCE TEST (BY GRADES)

Variance	Grade 3			Grade 4			Grade 5			Grade 6		
	d.f.	Sum of Squares	Mean Square	d.f.	Sum of Squares	Mean Square	d.f.	Sum of Squares	Mean Square	d.f.	Sum of Squares	Mean Square
Due to Practice Effect.	1	1,070.23	1,070.23	1	1,330.87	1,330.87	1	3,742.95	3,742.95	1	1,644.52	1,644.52
Between Individuals...	97	17,613.84	181.59	111	35,557.55	320.34	135	35,857.94	265.61	107	25,248.33	235.97
Error.....	97	1,636.77	16.87	111	1,636.13	14.74	135	2,272.55	16.83	107	1,462.48	13.67
Total.....	195	20,320.84	——	223	38,524.55	——	271	41,873.44	——	215	28,355.33	——

practice effect on this test is significantly greater in Grade 5 than in the other grades.

In cases such as this, where significant differences between grades or other groups are found, it is clear that it is wrong to combine the results and give only one analysis. The differences are often more useful than the similarities in determining the usefulness or appropriateness of a test. It is suggested, therefore, that in all cases an analysis similar to that shown in Table VII should be made. If significant differences between the grades or groups are found, the results should be shown separately for each group and not combined into a single analysis. If, in spite of these differences, a combined analysis is given, it should be clearly stated that the results apply only to the total of the groups and not necessarily to the component groups which form the total. The question of whether or not a useful interpretation of the results of the total analysis is possible can be answered only by an examination of the nature of the problem under consideration.

Note—

The values given in Table VII may be obtained by the method discussed in the first part of this section, or from the values given in Table IV. Using the results given in Table IV, we find for Grade 3, for example:

(1) *Sum of Squares corresponding to Practice Effect*

$$= \frac{(2{,}094)^2+(2{,}552)^2}{98} - \frac{(2{,}094+2{,}552)^2}{196}$$

$$= 111{,}199.39 - 110{,}129.16 = 1{,}070.23.$$

(2) *Sum of Squares corresponding to Between Individuals*

$$= \tfrac{1}{2}\left[53{,}408+77{,}042+2(62{,}518) - \frac{(2094+2552)^2}{98}\right]$$

$$= \tfrac{1}{2}\,[255{,}486 - 220{,}258.33]$$

$$= 17{,}613.84.$$

(3) *Sum of Squares corresponding to Error*

$$= \tfrac{1}{2}\left[53{,}408+77{,}042-2(62{,}518) - \frac{(2552-2094)^2}{98}\right]$$

$$= \tfrac{1}{2}\,[5{,}414 - 2{,}140.45]$$

$$= 1{,}636.77.$$

(4) *Sum of Squares corresponding to Total*

$$= 53{,}408+77{,}042 - \frac{(2094+2552)^2}{196}$$

$$= 130{,}450 - 110{,}129.16$$

$$= 20{,}320.84.$$

The corresponding values for the other grades may be calculated in a similar manner.

(3) *Comparison of the Accuracy of Physical and Mental Measurements*

Before we proceed to the next chapter, it is convenient at this stage to discuss the results of a little experiment which we carried out to compare the accuracy of physical and mental measurements. Since it was obviously more difficult to control an experiment involving mental measurements, it was decided to arrange an experiment with physical measurements which would correspond to the conditions generally found in measuring with mental tests. We had, therefore, to arrange a series of objects of different magnitudes and make two measurements of each. We could not, of course, arrange for the objects to change while being measured, or change between measurements, but otherwise the conditions seem to be comparable.

We chose to measure the lengths of strips of cardboard; 100 strips of different lengths were used. These were arranged so that we had a normal distribution of lengths; the distribution is shown in Table VIII (a class-interval 3/4 of an inch in length is used in this table).

TABLE VIII

DISTRIBUTION OF LENGTHS OF 100 STRIPS OF CARDBOARD (UNITS OF $\frac{1}{32}$ OF AN INCH)

Class Interval	Frequency
91–114	1
115–138	1
139–162	2
163–186	6
187–210	9
211–234	10
235–258	13
259–282	14
283–306	14
307–330	10
331–354	9
355–378	6
379–402	3
403–426	1
427–450	1
Total.....	100

4

This gave us a distribution of lengths for our physical measurements comparable to the distribution of ability for mental measurements.

The next step was to obtain two measurements of the length of each strip of cardboard in order that we might calculate a reliability coefficient comparable to those obtained in mental measurements. Somewhat to our surprise, considerable difficulty was experienced in reducing the accuracy of our measurements to the level found in the mental field. If we used an ordinary rigid measuring instrument, such as a ruler graduated in inches, the reliability coefficients were of the order of 0.99. It was obvious, therefore, that we had to use some non-rigid measuring instrument and deliberately introduce random errors of measurement. The plan which we finally adopted, after the trial and rejection of numerous others, is explained below.

A strip of rubber approximately one-half an inch in width and 16 inches long was cut from an inner tube of an automobile tire. This was stretched to twice its ordinary length and a scale marked on it (a unit on the scale at this tension corresponded approximately to one-eighth of an inch). These units were, of course, purely arbitrary, but this was immaterial. Finally, we fastened two clips firmly on the ends of this strip in order to make certain that the same length of rubber was used each time. The random errors of measurement were introduced by varying the tension each time the "rubber ruler" was used; as the clips were such that they could be slipped over the head of an ordinary nail, we could control the tension applied. We used six different tensions; these are denoted, from the highest to the lowest, by the numbers 1, 2, 3, 4, 5 and 6 in the following discussion.

In order to simplify the measuring process, and to control all extraneous factors, we proceeded as follows: A board approximately 10 inches in width and 40 inches in length was procured, and in this 7 nails were driven in the positions indicated in Figure 2. By slipping the clip at one end of the "rubber ruler" over the nail at 0, and the clip at the other end over the appropriate nail at the right hand side of the board, we could obtain any desired tension. By placing the strip of cardboard to be measured in position, a measurement of its length could be obtained with little difficulty. This procedure was followed throughout the experiment. (It should be noted that the scale as marked did not extend the full length of the "rubber ruler").

The strips of cardboard were arranged in random order of length and numbered from 1 to 100. Finally, a random series of tensions were chosen and two measurements of each strip of cardboard made according to this series. The figures given in Table IX show the number of each strip of cardboard, the tensions used in the measurements, and

TABLE IX

MEASUREMENTS OF THE LENGTHS OF 100 STRIPS OF CARDBOARD (USING A "RUBBER RULER")

Number of Strip	Tension Position		Length		Number of Strip	Tension Position		Length		Number of Strip	Tension Position		Length	
	1st Trial	2nd Trial	1st Trial	2nd Trial		1st Trial	2nd Trial	1st Trial	2nd Trial		1st Trial	2nd Trial	1st Trial	2nd Trial
1	1	2	58	62	35	5	1	81	65	69	3	1	90	81
2	2	1	57	54	36	6	6	58	58	70	4	2	43	39
3	3	3	70	70	37	1	2	49	52	71	5	4	65	61
4	4	6	77	87	38	2	5	55	65	72	6	5	78	74
5	5	2	50	42	39	3	1	36	32	73	1	1	49	49
6	6	4	71	63	40	4	5	68	72	74	2	3	45	47
7	1	3	60	66	41	5	4	53	50	75	3	2	82	75
8	2	1	36	34	42	6	3	64	54	76	4	6	55	62
9	3	6	59	70	43	1	5	47	59	77	5	4	67	63
10	4	3	73	69	44	2	1	64	61	78	6	5	72	68
11	5	5	94	94	45	3	3	72	72	79	1	4	58	68
12	6	4	50	44	46	4	5	80	84	80	2	3	53	56
13	1	2	63	67	47	5	6	97	103	81	3	6	80	96
14	2	1	60	57	48	6	2	102	81	82	4	3	56	53
15	3	6	61	73	49	1	4	55	65	83	5	5	86	86
16	4	1	63	54	50	2	3	63	67	84	6	1	73	55
17	5	6	62	66	51	3	4	84	89	85	1	5	70	87
18	6	3	59	49	52	4	3	43	40	86	2	2	87	87
19	1	4	44	51	53	5	2	56	47	87	3	6	87	105
20	2	2	55	55	54	6	4	76	67	88	4	5	60	64
21	3	3	66	66	55	1	6	64	85	89	5	3	51	45
22	4	5	72	77	56	2	5	78	93	90	6	2	91	72
23	5	6	56	59	57	3	1	79	71	91	1	5	57	72
24	6	4	79	69	58	4	2	74	66	92	2	1	73	69
25	1	1	56	56	59	5	6	77	83	93	3	5	31	35
26	2	4	59	66	60	6	3	47	39	94	4	6	26	30
27	3	3	48	48	61	1	1	65	65	95	5	2	89	75
28	4	1	81	69	62	2	5	56	67	96	6	4	55	49
29	5	2	84	71	63	3	4	78	82	97	1	3	42	46
30	6	6	67	67	64	4	2	71	64	98	2	6	31	40
31	1	5	85	106	65	5	3	76	68	99	3	1	57	51
32	2	4	49	55	66	6	6	94	94	100	4	4	78	78
33	3	2	98	93	67	1	2	38	41					
34	4	5	55	58	68	2	4	35	39					

the estimates of the length of the strip obtained by the use of the "rubber ruler". From the values given in the last two columns of the table, we may calculate an estimate of the reliability coefficient, or

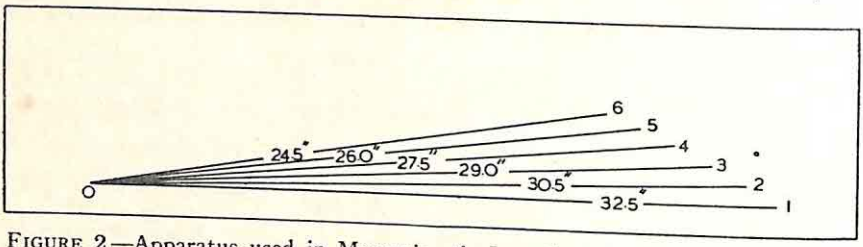

FIGURE 2.—Apparatus used in Measuring the Length of the Strips of Cardboard.

estimates of the absolute and relative accuracy of the measurements by using the method known as the analysis of variance.

Denoting by X_i and Y_i the values obtained in the first and second measurement, respectively, of the i-th strip of cardboard, and summation over all 100 values of i by Σ, we find:

$$\Sigma X_i = 6449$$
$$\Sigma Y_i = 6465$$
$$\Sigma X_i^2 = 442539$$
$$\Sigma Y_i^2 = 445847$$
$$\Sigma X_i Y_i = 440620$$

Using these results we have $r = 0.869$ as the estimate of reliability coefficient. Our measurements, therefore, are as reliable as those found in using a good intelligence test, or any other mental measuring instrument.

Using the analysis of variance method, we find the results shown in Table X. The differences between trials is not significant, but the variance between the strips is significantly greater than the error variance. We conclude, therefore, that our "rubber ruler" measures with sufficient accuracy to enable us to distinguish between the lengths of the strips of cardboard.

TABLE X

ANALYSIS OF VARIANCE OF THE MEASUREMENTS OF THE LENGTHS OF 100 STRIPS OF CARDBOARD

Variance	Degrees of Freedom	Sum of Squares	Mean Square
Between Trials...	1	1.28	1.28
Between Strips ...	99	50,956.02	514.71
Error.....	99	3,571.72	36.08
Total.....	199	54,529.02	———

Calculating an estimate of the sensitivity of our measuring instrument by the method explained earlier in this chapter, we find $\gamma = 2.6$.

Finally, we may use the square root of the error mean square as an estimate of the standard error of our measurements; in this case S_E is of the order of 6 units.

If these results are compared with those given in Tables I and II for the two forms of an intelligence test, it will be seen that, considered merely as a measuring instrument, our "rubber ruler" compares quite favourably with an intelligence test. It is realized, of course, that the conditions underlying the problems are not exactly the same, but it is felt that they are so nearly the same that the comparison is valid. This little experiment with a "rubber ruler", therefore, gives us a comparison of the accuracy of physical and mental measurements and, at the same time, gives us a clearer idea of the kind and magnitude of the errors which we make when we use a mental test as an instrument for measuring the ability of an individual, or of groups of individuals.

CHAPTER IV

EXPERIMENTAL RESULTS

Experiment I. *Estimation of Reliability Coefficients by the Split-half or Odds-even Method.*

It has been found that the estimates of the reliability of a test obtained by the use of different methods do not agree. The split-half estimate, generally obtained by correlating the scores made by the individuals on the odd and even items of the test, may be higher or lower than the estimate obtained by the use of comparable forms of the same test, or by the test-retest method. The purpose of the present section is to examine in detail the relationship between the split-half and comparable forms or test-retest estimates. The results given below refer to the relationship between the split-half and comparable forms estimates, but they apply equally well to the relationship between the split-half and test-retest estimates.

The theoretical relationship between these estimates is not difficult to determine and is well known. It will be re-developed here, however, since it is necessary to state clearly the assumptions made in determining the relationship, and to test the validity of these assumptions.

Denote by

Z_{1t} :—the score of the t-th individual on the first form of the test;

Z_{2t} :—the score of the t-th individual on the second form of the test;

X_{1t}, Y_{1t}:—the scores made by the t-th individual on the odd and even items, respectively, of the first test;

X_{2t}, Y_{2t}:—the scores made by the t-th individual on the odd and even items, respectively, of the second test;

Σ :—summation;

S :—the standard deviation, e.g.

$$S_{Z_1} = \sqrt{\frac{1}{N}\left\{\Sigma Z_{1t}^2 - \frac{(\Sigma Z_{1t})^2}{N}\right\}}$$

r :—the Pearson product-moment correlation coefficient;

N :—the number of individuals tested.

It may easily be shown that

$$S_{Z_1}^2 = S_{X_1}^2 + S_{Y_1}^2 + 2r_{X_1Y_1}S_{X_1}S_{Y_1} \quad \ldots\ldots\ldots (10)$$

$$S_{Z_2}^2 = S_{X_2}^2 + S_{Y_2}^2 + 2r_{X_2Y_2}S_{X_2}S_{Y_2} \quad \ldots\ldots\ldots (11)$$

$$r_{Z_1Z_2}S_{Z_1}S_{Z_2} = r_{X_1X_2}S_{X_1}S_{X_2} + r_{X_1Y_2}S_{X_1}S_{Y_2} \quad \ldots\ldots\ldots (12)$$
$$+ r_{Y_1X_2}S_{Y_1}S_{X_2} + r_{Y_1Y_2}S_{Y_1}S_{Y_2}$$

$$r_{Z_1Z_2} = \frac{r_{X_1X_2}S_{X_1}S_{X_2} + r_{X_1Y_2}S_{X_1}S_{Y_2} + r_{Y_1X_2}S_{Y_1}S_{X_2} + r_{Y_1Y_2}S_{Y_1}S_{Y_2}}{\sqrt{\{S_{X_1}^2 + S_{Y_1}^2 + 2r_{X_1Y_1}S_{X_1}S_{Y_1}\}\{S_{X_2}^2 + S_{Y_2}^2 + 2r_{X_2Y_2}S_{X_2}S_{Y_2}\}}} \quad \ldots\ldots (13)$$

Obviously $r_{Z_1Z_2}$ is determined exactly by the values of the six intercorrelations $r_{X_1X_2}$, $r_{X_1Y_2}$, $r_{Y_1X_2}$, $r_{Y_1Y_2}$, $r_{X_1Y_1}$, $r_{X_2Y_2}$, and the values of the four standard deviations S_{X_1}, S_{X_2}, S_{Y_1}, S_{Y_2}. If we assume that

$$\left.\begin{array}{l} (1)\ r_{X_1X_2} = r_{X_1Y_2} = r_{Y_1X_2} = r_{Y_1Y_2} = r_{X_1Y_1} = r_{X_2Y_2} = r, \text{ say} \\ \text{and } (2)\ S_{X_1} = S_{X_2} = S_{Y_1} = S_{Y_2} \end{array}\right\} \quad \ldots\ldots\ldots (14)$$

then equation (13) reduces to

$$r_{Z_1Z_2} = \frac{2r}{1+r} \quad \ldots\ldots\ldots (15)$$

Equation (15) is the Spearman-Brown formula used in determining the reliability of the whole test from that of the half-test, i.e.

$$r_w = \frac{2r_{\frac{1}{2}}}{1+r_{\frac{1}{2}}} \quad \ldots\ldots\ldots (16)$$

where r_w denotes the reliability coefficient for the whole test and $r_{\frac{1}{2}}$ denotes the reliability coefficient for the half-test.

In using the split-half method of estimating the reliability of the test, we use $r_{\frac{1}{2}} = r_{X_1Y_1}$ (or $r_{\frac{1}{2}} = r_{X_2Y_2}$) and substitute this value in equation (16) to obtain the estimate of the reliability of the test. In so doing, we make the assumptions shown above in equation (14) and also the implicit assumption that $r_{X_1Y_1}$, or $r_{X_2Y_2}$ as the case may be, is an unbiased estimate of the common r of equation (14). Experience has shown that the assumption of equal standard deviations, i.e. $S_{X_1} = S_{X_2} = S_{Y_1} = S_{Y_2} = S$, is generally satisfied in practice, but, of course, the validity of this assumption should be tested in each case. On the other hand, experience has also shown that the six intercorrelations are seldom, if ever, equal and that in particular $r_{X_1Y_1}$ or $r_{X_2Y_2}$ may be biased estimates of the assumed common value. The following example relates to the results for a small group of 56 pupils on two forms of an intelligence test, and shows clearly the kind of results which may be obtained. We found:

$r_{X_1X_2}=0.585$ $\quad S_{X_1}=4.523$

$r_{Y_1X_2}=0.652$ $\quad S_{Y_1}=4.247$

$r_{X_1Y_2}=0.707$ $\quad S_{X_2}=4.677$

$r_{Y_1Y_2}=0.743$ $\quad S_{Y_2}=5.039$

$r_{X_1Y_1}=0.766$ $\quad r_{Z_1Z_2}=0.765$

$r_{X_2Y_2}=0.772$

The standard deviations are all of the same order of magnitude, hence the assumption of equal standard deviations may be accepted as valid. If we assume only that the standard deviations are equal, equation (13) reduces to

$$r_{Z_1Z_2}=\frac{r_{X_1X_2}+r_{X_1Y_2}+r_{Y_1X_2}+r_{Y_1Y_2}}{2\sqrt{(1+r_{X_1Y_1})(1+r_{X_2Y_2})}} \qquad \cdots\cdots\cdots(17)$$

Substituting the values of the six correlation coefficients in (17), we find $r_{Z_1Z_2}=0.760$, which is an additional demonstration that the difficulty does not lie in the assumption regarding the equality of the standard deviations. Using the values of $r_{X_1Y_1}$ and $r_{X_2Y_2}$ in equation (16), however, we find:

for $r_{\frac{1}{4}}=r_{X_1Y_1}=0.766$, $\quad r_{Z_1Z_2}=0.867$

for $r_{\frac{1}{2}}=r_{X_2Y_2}=0.772$, $\quad r_{Z_1Z_2}=0.871$

which differ considerably from the observed value of 0.765.

The error, in this case at least, lies in using $r_{X_1Y_1}$ or $r_{X_2Y_2}$ as an estimate of the common coefficient of correlation of equation (14). Let us examine more closely the procedure underlying the determination of $r_{X_1Y_1}$ (the position is the same, of course, for $r_{X_2Y_2}$). The values of X_{1t} and Y_{1t} are determined, as explained above, from the scores on the odd and even items of the test. The formula generally used[1] in calculating $r_{X_1Y_1}$ is

$$r_{X_1Y_1}=\frac{\Sigma(X_{1t}Y_{1t})-\dfrac{(\Sigma X_{1t})\ (\Sigma Y_{1t})}{N}}{N\,S_{X_1}S_{Y_1}} \qquad \cdots\cdots\cdots(18)$$

We can see more clearly where the difficulty lies, however, if we write equation (18) in the following form:

$$r_{X_1Y_1}=\frac{S^2_{(X_1+Y_1)}-S^2_{(X_1-Y_1)}}{4S_{X_1}S_{Y_1}} \qquad \cdots\cdots\cdots(19)$$

where

$$S^2_{(X_1+Y_1)}=S^2_{Z_1} \qquad \cdots\cdots\cdots(20)$$

$$S^2_{(X_1-Y_1)}=\frac{1}{N}\left[\Sigma(X_{1t}-Y_{1t})^2-\frac{\{\Sigma(X_{1t}-Y_{1t})\}^2}{N}\right] \qquad \cdots\cdots\cdots(21)$$

[1]See Appendix A for a discussion of the procedure to be followed in estimating reliability coefficients in such cases.

The differences $X_{1t} - Y_{1t}$ are not independent of the total scores, in fact for large and for small total scores these differences must of necessity be small. This means, therefore, that when the proportion of small or large total scores in the sample is increased $S^2_{(X_1-Y_1)}$ must become smaller. At the same time, of course, we may increase the other term, $S^2_{(X_1+Y_1)}$, in the numerator of equation (19) so the net result will be a spurious increase in the value of $r_{X_1Y_1}$. This element of "spuriousness" will generally be present in the correlation of the scores on the odd and even items of a test and seems to be an inherent weakness of the method. The magnitude of the spurious element will always depend on the distribution of total scores on the test and in some cases it may not be important.

We tried to develop a correction term to allow for the bias involved, but we were not successful. The nature of the relationship between $S^2_{(X_1-Y_1)}$ and the total scores on a test is not difficult to determine. The results for three different types of distributions of total scores are shown in Tables XI, XII and XIII. In each of the three cases, of course, we kept the first term, $S^2_{(X_1+Y_1)}$, in the numerator of equation (19) constant.

In the first case, we chose a rectangular distribution of total scores, actually 5 papers for each total score from 6 to 60 inclusive (the total number of items on the test used was 75). Each group of five consecutive scores (i.e. 25 papers) was then treated as a unit, the scores made by each individual on the odd and even items found and the value of $S^2_{(X-Y_1)}$ calculated for each such unit. The results are shown in Table XI.

TABLE XI

VALUES OF $S^2_{(X_1-Y_1)}$ FOR VARIOUS TOTAL SCORE GROUPS: RECTANGULAR DISTRIBUTIONS, 25 PAPERS IN EACH GROUP

Total Score Groups	Values of $S^2_{(X_1-Y_1)}$
6–10	7.6
11–15	10.2
16–20	13.4
21–25	14.8
26–30	16.6
31–35	14.8
36–40	12.5
41–45	13.4
46–50	10.5
51–55	7.8
56–60	4.6

These results show a definite relationship between $S^2_{(X_1-Y_1)}$ and the total score; for larger and smaller scores than those considered $S^2_{(X_1-Y_1)}$ will continue to decrease, reaching the minimum value of zero for total scores of 0 and 75.

In practice we do not find rectangular distributions of scores like those in the example considered above. For this reason, therefore, we chose normal distributions of total scores in each unit for the following two examples. The results are shown in Tables XII and XIII; in the first case the distributions did not overlap, but in the second they did. We used 100 papers in each group of total scores for these cases, not 25 as in the previous case.

TABLE XII

VALUES OF $S^2_{(X_1-Y_1)}$ FOR VARIOUS TOTAL SCORE GROUPS: NORMAL DISTRIBUTIONS, NO OVERLAP, 100 PAPERS IN EACH GROUP

Total Score Groups	Values of $S^2_{(X_1-Y_1)}$
6–20	10.1
21–35	14.3
36–50	12.4
49–63	8.4

TABLE XIII

VALUES OF $S^2_{(X_1-Y_1)}$ FOR VARIOUS TOTAL SCORE GROUPS: NORMAL DISTRIBUTIONS, OVERLAPPING, 100 PAPERS IN EACH GROUP

Total Score Groups	Values of $S^2_{(X_1-Y_1)}$
1–35	12.4
18–52	14.4
41–75	8.1

The results here are similar to those shown in Table XI, but the differences are not as great. It is clear, however, that this effect will influence our estimates of the reliability coefficients, and in particular our estimates of the standard error of measurement. It follows, therefore, that we cannot use the analysis of variance method, which assumes that $S_{(X_1-Y_1)}$ is independent of the scores, in analysing data of this kind. No matter what method we use in analysing data of this kind, great care must be taken in the interpretation of the

results. It is doubtful if the split-half method can be used with any degree of confidence in estimating reliability coefficients. The estimates so obtained may be used as measures of "internal consistency" but not in all cases as measures of "reliability".

The results of two other experiments, which will now be considered, indicate that there are other factors, especially the test content, influencing the split-half estimates. It seems that no generalization can be made for all tests; each test must be considered separately.

Note:

Formula (19) may be used in the calculation of any correlation coefficient, and applies, with a slight change of notation, to the estimation of the reliability coefficient by the test-retest and comparable forms methods. It shows clearly the effect of selection of the group tested on the estimate of the reliability coefficient. Since the second term in the numerator is practically constant for a particular test, we may even find negative values of $r_{X_1Y_1}$ if the differences between the individuals tested are very small.

Experiment II. *Comparison of Comparable Forms, Test-Retest and Split-half Estimates of Reliability.*

Although many comparisons of the different estimates of reliability have been reported in the literature,[2] the experiments were not designed for the specific purpose we had in mind. We wished to compare the different estimates of reliability and at the same time to vary the length of time elapsing between the tests in order to determine what, if any, influence this had on the results. The Advanced Dominion Group Test of Intelligence[3] was chosen for this experiment; two comparable forms, *A* and *B*, were available, each consisting of 75 items of varying degrees of difficulty. For the test-retest experiments, Form *B* only was used. The tests were given to pupils in grades 9, 10, 11 and 12 but in the results given below no allowance has been made for between-grades differences as we were not interested in this particular factor. Each pupil was given either the two forms of the test or the same form twice, and the length of time elapsing between tests varied from a few minutes to 24 hours. The plan of the experiment is shown in Table XIV. It was impossible to obtain

[2]See Chapter I.

[3]Published by the Department of Educational Research, University of Toronto.

TABLE XIV

PLAN OF EXPERIMENT

Time tests were given	Number of Cases in	
	Test-retest group	Comparable forms group
In consecutive periods.....	146	86
In morning and afternoon of same day..	214	100
In corresponding periods of consecutive days.....	189	249

equal numbers of pupils for each group, but this does not matter much.

For each group we have, using the usual correlation technique, two split-half estimates and a test-retest or comparable forms estimate of the reliability of the test. These estimates are given in Table XV. The test-retest estimates are consistently higher than the comparable forms estimates, and the split-half estimates are generally higher than the comparable forms and lower than the test-retest estimates. An interesting point is the increase in the split-half estimates from the first to the second form for both test-retest and comparable forms

TABLE XV

COMPARISON OF TEST-RETEST, COMPARABLE FORMS AND SPLIT-HALF ESTIMATES OF RELIABILITY

Time tests were given	Reliability Coefficients					
	Test-retest Form B	Split-half		Comparable Forms Forms A & B	Split-half	
		Form B (First)	Form B (Second)		First Form (A)	Second Form (B)
Consecutive periods.....	0.937	0.880	0.912	0.839	0.833	0.889
Morning and afternoon of same day...	0.914	0.882	0.909	0.881	0.921	0.910
Consecutive days.....	0.932	0.889	0.923	0.887	0.904	0.923

groups. A similar change has been noted in other cases, but very often we find a decrease instead of an increase. This effect seems to be caused by practice lengthening or shortening the effectual length[4] of the test. If the test is rather easy in the first place, we generally find a decrease; if the test is a little too difficult, we tend to find an increase.

If we use the analysis of variance method in analysing these data, we find a simple explanation of the difference between the test-retest and comparable forms estimates. The complete results of this analysis are presented in Table XVI. The values are arranged to assist in the comparison of results for each group on the two experimental methods, and between groups for each experimental method.

The error variance seems to be independent of the length of time elapsing between the tests. In each case, however, the test-retest and comparable forms estimates of error are significantly different; the comparable forms experimental method yields consistently higher estimates. The tests of significance of the differences are made as shown below [see 53].

1. *Consecutive periods*

   Calculate $F=\frac{15.5}{9.1}=1.7$ and refer to Snedecor's tables of F with degrees of freedom $n_1=85$ and $n_2=145$.

2. *Morning and afternoon of same day*

   Calculate $F=\frac{14.7}{10.3}=1.4$ and refer to Snedecor's tables of F with degrees of freedom $n_1=99$ and $n_2=213$.

3. *Consecutive days*

   Calculate $F=\frac{15.9}{10.6}=1.5$ and refer to Snedecor's tables of F with degrees of freedom $n_1=248$ and $n_2=188$.

In all cases the differences are significant, so we may conclude that different experimental methods give different estimates of the errors of measurement.

[4]By this is meant the number of items which are used by the individuals, not necessarily the number of items on the test. It is interesting to note that we have no exact measure of the "true" or "effectual" length of a test, at least not as far as the authors are aware. The reader can easily convince himself that the number of items composing the test is a poor measure of its length by considering a simple case. For very easy or very difficult tests, clearly the discrimination between individuals is obtained on relatively few items, the remaining items being useless as far as the purpose of the test is concerned. In these cases the number of items composing the test may bear little or no relation to its "true" or "effectual" length.

TABLE XVI

COMPARISON OF RESULTS OF THE ANALYSIS OF VARIANCE OF DATA OBTAINED BY USING TEST-RETEST AND COMPARABLE FORMS EXPERIMENTAL METHODS

Time tests were given	Variance	Test–Retest			Comparable Forms		
		Degrees of Freedom	Sum of Squares	Mean Square	Degrees of Freedom	Sum of Squares	Mean Square
Consecutive periods	Between Trials	1	3,563	3,563	1	1,856	1,856
	Between Individuals	145	36,154	249	85	14,368	169
	Error	145	1,335	9.1	85	1,321	15.5
	Total	291	41,052	—	171	17,545	—
Morning and afternoon of same day	Between Trials	1	3,587	3,587	1	1,480	1,480
	Between Individuals	213	45,596	214	99	22,591	228
	Error	213	2,187	10.3	99	1,457	14.7
	Total	427	51,370	—	199	25,528	—
Consecutive days	Between Trials	1	6,638	6,638	1	953	953
	Between Individuals	188	52,055	277	248	65,325	263
	Error	188	1,994	10.6	248	3,940	15.9
	Total	377	60,687	—	497	70,218	—

It is difficult to determine the cause of this effect. It may be due to the memory factor entering in the test-retest method but it is more likely that, in spite of the care taken in constructing the test, the two forms were not exactly comparable.

With regard to the other effects, practice appears to be more important in the test-retest method—which may account for part of the

difference discussed in the preceding paragraph. We note also that for some reason the first comparable forms group was more homogeneous than the others. This does not affect the between trials or error estimates for this group, of course, but it does affect the estimate of the reliability coefficient and hence explains the lower values shown for this group in Table XV.

On the basis of all these results, we may conclude that the estimates of the reliability of a test obtained by the use of different experimental methods are not exactly comparable. In reporting on the reliability of a test we should, therefore, state which experimental method was used and if possible give results for each method. It is clear that we cannot compare the reliability of different tests unless complete and detailed information on these points is available.

Experiment III. *Comparison of Test-Retest and Split-half Estimates of Reliability for a Battery of Sub-tests.*

Tests may be composed of a large number of items, of the same or different content, or of a series of short sub-tests of different content. In the above two experiments we used tests of the first type but for this experiment we chose a test composed of six relatively short sub-tests of different content. We wished to determine the relationship between the test-retest and split-half estimates of reliability separately for each sub-test and for the whole test and, in addition, the effect of varying the time between tests on these estimates and their relationship. These estimates have also been compared with the estimates given by the application of Kuder and Richardson's formula (20) [63] for two of the groups of children considered. Other questions relating to the reliability of a battery of tests or sub-tests have been considered in another section of the bulletin.

The test chosen for use in this experiment was the Revised Beta Examination prepared by C. E. Kellogg and N. W. Morton of McGill University, Montreal, Canada. The test is composed of six short sub-tests (also six exercises, one for each sub-test) and the material is non-verbal in content. The content of each sub-test and maximum score are shown in Table XVII; the total score is obtained by adding the unweighted sub-test scores.

The group of children tested consisted of 5 classes of Grade IX pupils in an Ontario school. Altogether 175 pupils were tested, but only 156 of these were present for both the test and retest. The interval elapsing between the tests was one-half day, 1 day, 3 days, 1 week and 5 weeks for the classes designated A, B, E, D and C, respectively. The

TABLE XVII

CONTENT AND MAXIMUM SCORE OF EACH SUB-TEST OF REVISED BETA EXAMINATION

Sub-test	Content	Maximum Score
1	Maze (10 mazes)	10
2	Digit symbol (90 items)	30
3	Common-sense picture discrimination (picture absurdities) (20 items)	20
4	Form board (18 items)	18
5	Picture completion (20 items)	20
6	Number checking (50 items)	25

classes, testing dates, number of pupils, mean scores and standard deviations for each sub-test are shown in Table XVIII. The tests seem to be too easy for Grade IX pupils and, partly for this reason, the spread of scores is not large (for the whole test, the standard deviations varied from 6.52 to 9.90). As homogeneous grouping is not used in this school, there is very little difference between the classes.

There is, unfortunately, no general pattern evident in the results except for the practice effect, but even this varies considerably. The standard deviations on the second trial are both larger and smaller than those on the first trial. Except for those cases in which the average score on the second trial was very high compared with the number of items on the test, e.g. in the case of sub-test 6 in Class C, it is impossible to determine just what is the net effect of practice. In some cases it seems to increase, and in others decrease, the "effectual" length of the test. As far as determining a measure of the true length of the test is concerned, these results indicate that some function of the standard deviation should be considered.

The position is much clearer when we consider the comparison of the test-retest and split-half estimates of reliability. The results are fairly consistent for each sub-test but they vary from one sub-test to another. For this reason, therefore, the data given in Table XIX are arranged by sub-tests, not by classes as in Table XVIII.

Sub-tests 1, 2 and 6 form a group giving similar results. In these

TABLE XVIII

REVISED BETA EXAMINATION: MEAN SCORE AND STANDARD DEVIATION OF SCORES ON EACH SUB-TEST (BY CLASSES)

Class	Testing Dates	Number of Pupils	Sub-test	Mean Score		Standard Deviation	
				1st Trial	2nd Trial	1st Trial	2nd Trial
A	Dec. 15th A.M. Dec. 15th P.M.	32	1	7.7	8.0	1.41	1.09
			2	23.3	25.4	3.18	3.77
			3	13.3	14.2	2.99	3.11
			4	10.0	11.5	2.88	3.12
			5	14.8	16.3	2.45	2.11
			6	19.8	20.4	2.35	2.38
B	Dec. 14th Dec. 15th	34	1	7.6	8.5	1.19	1.22
			2	22.3	25.3	3.29	2.93
			3	12.8	14.4	2.26	2.46
			4	9.8	11.5	3.04	3.17
			5	14.8	16.5	2.74	2.43
			6	18.9	20.5	2.29	2.29
C	Dec. 15th Jan. 22nd	29	1	7.6	8.6	1.16	1.25
			2	21.1	24.4	3.08	3.78
			3	12.2	14.1	2.16	2.35
			4	10.6	12.2	3.23	3.44
			5	14.4	16.0	2.23	2.25
			6	18.2	23.0	2.55	1.64
D	Dec. 14th Dec. 21st	32	1	7.1	8.6	1.34	0.96
			2	21.3	25.4	3.33	3.20
			3	12.8	14.9	2.86	2.49
			4	11.0	12.6	2.96	2.55
			5	15.1	17.1	1.95	1.94
			6	20.5	21.0	1.89	2.16
E	Dec. 15th Dec. 18th	29	1	7.7	9.1	1.44	1.06
			2	23.7	26.6	4.85	3.03
			3	12.5	13.3	2.19	1.75
			4	10.7	12.6	3.13	3.24
			5	14.4	16.7	2.28	1.84
			6	20.6	22.2	2.53	2.15

cases the split-half estimates, except for two cases for sub-tests 1 and 6, are consistently higher than the test-retest estimates. The explanation of this difference seems to lie in the content of the sub-tests. In sub-test 2, for example, the items are all of the same difficulty and are

TABLE XIX

REVISED BETA EXAMINATION: COMPARISON OF TEST–RETEST AND SPLIT-HALF ESTIMATES OF RELIABILITY (BY SUB-TESTS)

Sub-test	Class	Reliability Coefficients		
		Test-Retest	Split-half*	
			1st Trial	2nd Trial
1	A	0.469	0.692	0.238
	B	0.363	0.407	0.467
	C	0.301	0.409	0.442
	D	0.498	0.714	0.410
	E	0.562	0.591	0.713
2	A	0.714	0.949	0.986
	B	0.467	0.967	0.947
	C	0.591	0.984	0.978
	D	0.842	0.974	0.983
	E	0.686	0.979	0.966
3	A	0.897	0.708	0.724
	B	0.713	0.591	0.825
	C	0.566	0.398	0.507
	D	0.676	0.590	0.567
	E	0.560	0.378	0.091
4	A	0.878	0.712	0.833
	B	0.789	0.758	0.813
	C	0.811	0.895	0.811
	D	0.826	0.796	0.560
	E	0.769	0.825	0.812
5	A	0.801	0.752	0.673
	B	0.868	0.828	0.680
	C	0.661	0.566	0.642
	D	0.727	0.502	0.477
	E	0.851	0.501	0.665
6	A	0.660	0.782	0.781
	B	0.539	0.779	0.790
	C	0.091	0.793	0.306
	D	0.848	0.533	0.817
	E	0.674	0.731	0.833
Total	A	0.920	0.916	0.915
	B	0.773	0.936	0.908
	C	0.623	0.876	0.848
	D	0.856	0.851	0.810
	E	0.841	0.855	0.847

*For whole test.

of such a nature that if a pupil gets one item correct, he is almost certain to obtain the correct answer for the adjoining item. A somewhat similar situation exists for sub-tests 1 and 6 but here the effect is not so marked. Clearly, for tests of such content the split-half method should not be used in estimating the reliability of the test.

Sub-tests 3 and 5 form another group but in this case the split-half estimates are consistently lower than the test-retest estimates (except for one value for sub-test 3). These sub-tests are similar in content; in 3 the pupil is asked to mark the absurd picture while in 5 he is asked to complete the drawing. Since the items in these sub-tests are not arranged properly in order of difficulty, the two halves of the test formed by the odd and even items are not exactly comparable. This effect may lower the split-half estimate. On the other hand, the items are of such a nature that the pupil is likely to remember the answer given on the first test. This effect would tend to increase the test-retest estimates. It is probable that some combination of these two effects accounts for the observed differences in the results.

In the case of sub-test 4, the test-retest and split-half estimates agree very well; the split-half estimates are higher in half the cases and lower in the other half. The items of this sub-test seem to be satisfactorily arranged in order of difficulty, except possibly item 7, and it is unlikely that the pupils could remember very well the answers given on the first testing.

A comparison, for two of the classes, of the test-retest, split-half and Kuder-Richardson estimates of reliability is shown in Table XX. The Kuder-Richardson estimates, which are really estimates of the internal consistency of the sub-tests, agree better with the split-half than with the test-retest estimates. This indicates that the split-half method gives estimates of the internal consistency of the test, which may be very different from the reliability of the test as discussed in the earlier sections of this bulletin. Clearly, the different methods give results which are not always comparable, and in some cases it is difficult to determine exactly what is measured.

A comparison, for the test as a whole, of the test-retest and split-half estimates of reliability is given in the last section of Table XIX. The estimates agree fairly well for three of the classes, A, D and E, but differ considerably for the remaining two. It will be realized, of course, that strange results may be obtained when we combine the unweighted scores for sub-tests of such different content.

The Revised Beta Examination seems to behave rather erratically. Considering only the test-retest coefficients, for example, the estimates

TABLE XX

REVISED BETA EXAMINATION: COMPARISON OF TEST-RETEST, SPLIT-HALF, AND KUDER-RICHARDSON ESTIMATES OF RELIABILITY (FOR 2 CLASSES)

Class	Sub-test	Reliability Coefficients				
		Test-Retest	Split-half		Kuder-Richardson	
			1st Trial	2nd Trial	1st Trial	2nd Trial
B	1	0.363	0.407	0.467	0.260	0.394
	2	0.467	0.967	0.947	0.943	0.912
	3	0.713	0.591	0.825	0.437	0.575
	4	0.789	0.758	0.813	0.743	0.755
	5	0.868	0.828	0.680	0.700	0.666
	6	0.539	0.779	0.790	0.640	0.607
E	1	0.562	0.591	0.713	0.568	0.499
	2	0.686	0.979	0.966	0.968	0.928
	3	0.560	0.378	0.091	0.374	0.011
	4	0.769	0.825	0.812	0.766	0.781
	5	0.851	0.501	0.655	0.566	0.565
	6	0.674	0.731	0.833	0.686	0.603

of reliability vary from 0.623 to 0.920. These differences might, of course, be due to differences in the variability of the individuals in different classes with respect to the ability measured by the test. The analysis of variance of these data given in Table XXI, however, shows that while there may be certain differences between the classes with regard to this factor, more significant differences occur in the estimates of the errors of measurement. The mean square for error varies from 7.79 in Class D to 28.98 in Class C, and these changes are not related to the length of time elapsing between the tests (the practice, or between trials, effect is affected by this factor). The samples are small,

TABLE XXI

REVISED BETA EXAMINATION: ANALYSIS OF VARIANCE OF TOTAL SCORES (BY CLASSES)

Variance	Class A			Class B			Class E			Class D			Class C		
	d.f.	Sum of Squares	Mean Square	d.f.	Sum of Squares	Mean Square	d.f.	Sum of Squares	Mean Square	d.f.	Sum of Squares	Mean Square	d.f.	Sum of Squares	Mean Square
Between Trials	1	784	784	1	1874.25	1874.25	1	1710.77	1710.77	1	2268.14	2268.14	1	3084.98	3084.98
Between Individuals.....	31	5902	190.39	33	4040.78	122.45	28	3598.48	128.52	31	2923.98	94.32	28	3486.48	124.52
Error.....	31	245	7.90	33	525.25	15.92	28	385.73	13.78	31	241.36	7.79	28	811.52	28.98
Total.....	63	6931	——	67	6440.28	——	57	5694.98	——	63	5433.48	——	57	7382.98	——

of course, but even taking this into account, it is surprising to find significant differences in estimates of the errors of measurement.

The conclusions which follow from these experimental results are summarized below:

(1) Different experimental methods of measuring reliability yield results which are not always comparable.

(2) The Kuder-Richardson and split-half methods give measures of the internal consistency of the test; this may or may not be the same thing as the reliability of the test.

(3) For certain experimental methods, the content of the test affects the estimates of reliability.

(4) Practice may lengthen or shorten the "true" length of the test, and hence affect the reliability coefficients.

(5) The length of time elapsing between tests seems to have little effect on the estimates of reliability (except for the practice effect noted above).

(6) The number of items composing a test is not a very efficient measure of the "true" or "effectual" length of a test. We need a better definition and measure of length than has as yet been proposed.

(7) The estimates of errors of measurement are generally, but not always, constant for a particular test. They also are affected by the particular experimental method employed.

CHAPTER V

THE ESTIMATION OF TEST RELIABILITY BY THE METHOD OF RATIONAL EQUIVALENCE

Among recent developments in the theory and estimation of test reliability is a method developed by G. F. Kuder and M. W.Richardson [63, 88] whereby the reliability of tests may be estimated, on the basis of certain assumptions, from item analysis parameters. This method, described by its authors as the method of rational equivalence, is developed on a foundation of fundamental test structure theory, which renders explicit the essential determiners of test functioning. The basic observation underlying the method is that both test variance and test reliability, relative to a defined population, are functions not only of the individual item variances and item reliabilities but also of the inter-item covariances. The inter-item covariances are in turn in part a function of the item difficulty values and the item content, two of the basic determiners of a test's internal consistency.

The term rational equivalence results from an operational definition of equivalence deduced from formulae for the correlation of sums. A test Z_1 is presumed to be equivalent to a hypothetical parallel form $Z_{1'}$ when every item i of Z_1 is interchangeable with a corresponding item i' of $Z_{1'}$, every pair of items being similar with respect to difficulty and content. The further assumption is that all corresponding correlations are equal. Thus rational equivalence is defined such that $i=i'$, and $r_{ij}=r_{i'j}=r_{i'j'}$. Relevant to the above is the observation that precisely similar assumptions underly the Spearman-Brown formula for estimating increased reliability with increased length of test. Somewhat broader assumptions, however, underly the Spearman-Brown formula when it is used to estimate reliability by augmenting the correlation between split-halves.

Now in the derivation of their formulae Kuder and Richardson made certain assumptions, over and above the equivalence condition, which if necessary for final proof would detract seriously from the practical usefulness of their method, since many of the conditions which they found it necessary to specify are rarely if ever attained in practice, or for that matter even roughly approximated. Furthermore a few of their assumptions are inconsistent one with the other.

Since several of these formulae are rapidly coming into use for the estimation of the reliability of tests, it is of some moment to examine these formulae carefully, and to re-derive several of them on a greater parsimony of specified condition. Indeed it may be shown that the Kuder and Richardson formula (20), which was found empirically to yield very satisfactory results, may be derived on the basis of the equivalence assumption only. In general we may state that these authors fell into the common error of specifying conditions that were sufficient but unnecessary.

We may write the intercorrelations between the n_1 test items of Z_1 and the n_1 assumed equivalent items of $Z_{1'}$ in the form of a pooling square [112], weighting each item according to its standard deviation, as follows:[5]

	S_1	S_2	S_3	. . .	S_{n_1}	$S_{1'}$	$S_{2'}$	$S_{3'}$	. . .	$S_{n'_1}$
S_1	1	r_{12}	r_{13}	. . .	r_{1n}	$r_{11'}$	$r_{12'}$	$r_{13'}$	. . .	$r_{1n'}$
S_2	r_{12}	1	r_{23}	. . .	r_{2n}	$r_{12'}$	$r_{22'}$	$r_{23'}$	. . .	$r_{2n'}$
S_3	r_{13}	r_{23}	1	. . .	r_{3n}	$r_{13'}$	$r_{23'}$	$r_{33'}$	. . .	$r_{3n'}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	. . .	. . .	. . .	. . .	$r_{n(n-1)}$	. . .	. . .	. . .	. . .	$r_{n'(n-1)}$
S_{n_1}	r_{1n}	r_{2n}	r_{3n}	$r_{n(n-1)}$	1	$r_{1n'}$	$r_{2n'}$	$r_{3n'}$	$r_{n'(n-1)}r_{nn'}$	
$S_{1'}$	$r_{1'1}$	$r_{1'2}$	$r_{1'3}$	. . .	$r_{1'n}$	1	$r_{1'2'}$	$r_{1'3'}$	. . .	$r_{1'n'}$
$S_{2'}$	$r_{1'2}$	$r_{2'2}$	$r_{2'3}$	. . .	$r_{2'n}$	$r_{1'2'}$	1	$r_{2'3'}$	. . .	$r_{2'n'}$
$S_{3'}$	$r_{1'3}$	$r_{2'3}$	$r_{3'3}$	. . .	$r_{3'n}$	$r_{1'3'}$	$r_{2'3'}$	1	. . .	$r_{3'n'}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	. . .	. . .	. . .	. . .	$r_{n(n'-1)}$	. . .	. . .	. . .	. . .	$r_{n'(n'-1)}$
$S_{n'_1}$	$r_{1'n}$	$r_{2'n}$	$r_{3'n}$	$r_{n(n'-1)}r_{n'n}$		$r_{1'n'}$	$r_{2'n'}$	$r_{3'n'}$	$r_{n'(n'-1)}$	1

[5]The standard deviation of a dichotomously scored test item, i, is given by the formula

$$s_i = \sqrt{p_i q_i}$$

where s_i = the standard deviation of item i.
p_i = the proportion of persons passing item i.
q_i = the proportion of persons failing item i.

The correlation between any two dichotomously scored test items is understood in the present paper to be given by

$$r_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i p_j q_i q_j}}$$

where p_{ij}=the proportion of persons passing both items i and j.

The equivalence assumption implies that in the above correlation matrix $r_{12}=r_{1'2'}=r_{1'2}$. The sum of the weighted elements in the upper left hand quadrant of the above matrix is the variance of Z_1, the sum of the weighted elements in the lower right-hand quadrant is the variance of $Z_{1'}$, and the sum of the weighted elements in either the upper right-hand quadrant or the lower left-hand quadrant is the covariance. Now, since the correlation between two tests is given by dividing the covariance by the square root of the product of the two variances, we may write

$$r_{tt} = \frac{2\sum_{i,j\ (i<j)} r_{ij'} S_i S_{j'} + \sum r_{ii'} S_i^2}{[\sum S_i^2 + 2\sum_{i,j\ (i<j)} r_{ij} S_i S_j]\,[\sum S_{i'}^2 + 2\sum_{i,j\ (i<j)} r_{i'j'} S_{i'} S_{j'}]}$$

where r_{tt} = reliability of test,
r_{ij} = correlation between items i and j,
S_i = standard deviation of item i,
S_j = standard deviation of item j,
$r_{ii'}$ = reliability of item i.

But on the basis of the equivalence assumption

$$r_{ij}=r_{i'j'}=r_{i'j}$$

and $$S_i=S_{i'};\ S_j=S_{j'}$$

consequently $$2\sum r_{ij'}S_iS_j=2\sum r_{i'j'}S_{i'}S_{j'}=2\sum r_{ij'}S_iS_{j'}$$

and $$\sum S_i^2=\sum S_{i'}^2$$

therefore $$r_{tt}=\frac{S_t^2-\sum S_i^2+\sum r_{ii'}S_i^2}{S_t^2} \qquad \ldots\ldots\ldots (22)$$

where S_t^2 is the variance of the test.

Formula (22) is the basic formula underlying Kuder and Richardson's discussion of test reliability.

Another interesting derivation of formula (22) is given here. The error variance of a single test item may be found by the formula

$$S_{ei}^2=S_i^2(1-r_{ii'}) \qquad \ldots\ldots\ldots (23)$$

where S_{ei}^2 is the error variance of item i. On the assumption that errors of measurement are uncorrelated, the error variances of the n items on a test may be summed to give the error variance of the whole test. Thus

$$S_e^2=\sum S_i^2-\sum r_{ii'}S_i^2 \qquad \ldots\ldots\ldots (24)$$

where S_e^2 is the error variance of the whole test.

But the usual formula for the error variance of a test score is

$$S_e^2 = S_t^2 - r_{tt} S_t^2$$

hence

$$r_{tt} = 1 - \frac{S_e^2}{S_t^2} \qquad \ldots\ldots\ldots (25)$$

Substituting (23) in (25) we obtain

$$r_{tt} = \frac{S_t^2 - \sum S_i^2 + \sum r_{ii'} S_i^2}{S_t^2}$$

which is identical with formula (22).

We could now by making certain assumptions estimate the term $\sum r_{ii'} S_i^2$, and derive the Kuder-Richardson formula (20). This formula, however, is capable of derivation by more direct means.

The variance of a test of n items, written as a function of the item variances and inter-item covariances, is as follows:

$$\begin{aligned} S_t^2 &= \sum S_i^2 + 2 \sum_{i,j\ (i<j)} r_{ij} S_i S_j \\ &= n \overline{S_i^2} + n(n-1)\, \overline{r_{ij} S_i S_j} \end{aligned} \qquad \ldots\ldots\ldots (26)$$

where

$\overline{S_i^2}$ = average item variance.

$\overline{r_{ij} S_i S_j}$ = average item covariance.

If we assume that there exists a hypothetical parallel form of the test, also composed of n items, then the variance of the sum of these two tests will be

$$S_T^2 = 2n \overline{S_i^2} + 2n(2n-1) \overline{r_{ij} S_i S_j} \qquad \ldots\ldots\ldots (27)$$

where S_T^2 is the variance of the sum of scores on the two equivalent forms. We know, however, from formulae for the correlation of sums that

$$S_T^2 = 2 S_t^2 (1 + r_{tt}) \qquad \ldots\ldots\ldots (28)$$

where r_{tt} is the reliability coefficient; that is, the correlation between a test and its hypothetical equivalent form.

Substituting the values of S_t^2 and S_T^2 from (26) and (27) respectively in (28), and solving for r_{tt}, we obtain

$$r_{tt} = \frac{n}{n-1} \cdot \frac{S_t^2 - \sum S_i^2}{S_t^2} \qquad \ldots\ldots\ldots (29)$$

This formula is identical with the Kuder-Richardson formula (20).

If we examine the assumptions in the above derivation, we see that they are not quite identical with the equivalence assumption as previously stated. The equivalence assumption specified that $r_{ij} = r_{i'j'} =$

$r_{i'j}$, and $S_i = S_{i'}$. Here we have made a somewhat less rigorous assumption, namely that $\overline{r_{ij}S_iS_j} = \overline{r_{i'j'}S_{i'}S_{j'}} = \overline{r_{i'j}S_{i'}S_j}$. Thus we have assumed that our covariances are *on the average* equal.

The average inter-item covariance of a test is known exactly, and is of interest as a statistic descriptive of a test's internal consistency. It may vary within the limits $-.25/n-1$ and .25. Similar observation indicates that the average inter-item correlation can never be less than $-1/n-1$. This observation applies not only to test items but also to test batteries. Hence when the number of items is large it is mathematically almost impossible to obtain a negative average inter-item correlation.

If there is reason to believe that all test items are of equal difficulty, then the term $\sum S_i^2$, which is of course $n\bar{p}_i\bar{q}_i$, will be equal to $n\bar{p}_i\bar{q}_i$, and formula (29) may be written

$$r_{tt} = \frac{n}{n-1} \cdot \frac{S_t^2 - n\bar{p}\bar{q}}{S_t^2} \qquad \ldots\ldots\ldots (30)$$

where

$$\bar{p} = \frac{M_t}{n} \qquad \ldots\ldots\ldots (31)$$

A number of formulae employing item-test correlation may be derived on the basis of the observation that

$$\sum r_{it} S_i S_t = S_t^2 \qquad \ldots\ldots\ldots (32)$$

Such formulae involve somewhat more arithmetical labour than formulae already given, but represent no improvement in the estimation of test reliability.

Jackson [52] has suggested that the accuracy of measurement of a test should be described in terms of a sensitivity statistic γ, defined as

$$\gamma = \sqrt{\frac{S_t^2 - S_e^2}{S_e^2}} \qquad \ldots\ldots\ldots (33)$$

Hence the sample value of γ is related to the reliability of a test by the formula

$$\gamma = \sqrt{\frac{r_{tt'}}{1 - r_{tt'}}} \qquad \ldots\ldots\ldots (34)$$

Substituting formula (29) in formula (34) we obtain a value of γ estimated by the method of rational equivalence; thus

$$\gamma = \sqrt{\frac{S_t^2(n-1)}{n\sum_{i=1}^{n} S_i^2 - S_t^2} - 1} \qquad \ldots\ldots\ldots (35)$$

The statistic γ thus calculated may be used in the determination of a probability of making an error of measurement by chance alone greater than or equal to $\sqrt{S_t^2 - S_e^2}$ units.

Formula (29) above is identical with the Kuder and Richardson formula (20) which those authors found on the basis of empirical evidence to yield very satisfactory results. Their derivation, however, required the assumption that the matrix of inter-item correlations have a rank of 1, and that all the inter-item correlations be equal. These assumptions while sufficient are unnecessary, and indeed if they were necessary the formula would be of little practical value since few tests approximate to these conditions.

As mentioned previously several of the formulae derived by Kuder and Richardson are based on assumptions that are incompatible one with the other. Their formula (14), for example,

$$r_{tt'} = \frac{S_t^2 - \sum_{i=1}^{n} S_i^2}{(\sum_{i=1}^{n} S_i)^2 - \sum_{i=1}^{n} S_i^2} \cdot \frac{(\Sigma S_i)^2}{S_t^2} \qquad \ldots\ldots\ldots (36)$$

is presumably derived on the assumption that the rank of the matrix of inter-item correlations is 1, and that all the intercorrelations are equal. The difficulty values of the items are allowed to vary over a wide range. Now if the items are homogeneous with respect to difficulty and content, then all the intercorrelations will be approximately equal, and the inter-item correlation matrix will have a rank of 1. If, however, the items are heterogeneous with respect to difficulty the intercorrelations will not be equal, since the correlation between two test items is not independent of their difficulties. In general the greater the difference in difficulty between two test items the smaller the correlation between them. Furthermore, if the items are heterogeneous with respect to difficulty, although homogeneous with respect to content, the matrix of correlations will not be of rank 1, since differences in difficulty are represented in the factorial configuration describing the matrix of inter-item correlations as additional factors. Hence we see that the assumption of rank 1 and equal intercorrelation is incompatible with the provision that the difficulty values of the items be allowed to vary over a wide range.

The Kuder-Richardson formulae furnish useful statistics in describing the properties of tests even although the equivalence assumption is not satisfied. Under such circumstances, however, we are not justified in describing the obtained coefficients as reliability coeffi-

cients, since implicit in the reliability concept is the idea of repeated measurement. If we are unwilling to make the equivalence assumption we may refer to the coefficients obtained by the Kuder-Richardson formulae as *consistency coefficients*; that is, coefficients descriptive of a test's internal consistency. Reliability coefficients are, therefore, identical with consistency coefficients when the equivalence condition is satisfied.

CHAPTER VI

BATTERY RELIABILITY

Battery Reliability

The reliability of test batteries may be conveniently calculated from formulae for the correlation of sums. If $z_1, z_2, \ldots . z_n$ are n initial tests and $z_{1'}, z_{2'}, \ldots . z_{n'}$ are corresponding alternative forms, then the correlation between the simple sum of scores on the initial tests and the sum of scores on the alternative forms may be written

$$R = \frac{\sum r_{ii'} S_i S_{i'} + 2 \sum\limits_{i,j'(i<j')} r_{ij'} S_i S_{j'}}{\sqrt{\left[\sum S_i^2 + 2 \sum\limits_{i,j\ (i<j)} r_{ij} S_i S_j\right]\ \left[\sum S_{i'}^2 + 2 \sum\limits_{i',j'\ (i'<j')} r_{i'j'} S_{i'} S_{j'}\right]}} \quad \ldots\ldots\ldots (37)$$

where $r_{ii'}$ = reliability of test z_i,
S_i = standard deviation of test z_i,
$S_{i'}$ = standard deviation of test $z_{i'}$.

If we can assume that $S_i = S_{i'}$ ..... (38)

and $r_{ij} = r_{i'j'} = r_{i'j} = r_{ij'}$ ..... (39)

then the reliability of the battery may be written

$$R = \frac{\sum r_{ii'} S_i^2 + 2 \sum\limits_{i,j\ (i<j)} r_{ij} S_i S_j}{\sum S_i^2 + 2 \sum\limits_{i,j\ (i<j)} r_{ij} S_i S_j} \quad \ldots\ldots\ldots (40)$$

Formulae (37) and (40) relate to the reliability of the simple sum of scores on n tests. If the scores on each test are expressed in standard measure formula (37) may be written as

$$R = \frac{\sum r_{ii'} + 2 \sum\limits_{i,j'\ (i<j')} r_{ij'}}{\sqrt{\left[n + 2 \sum\limits_{i,j\ (i<j)} r_{ij}\right]\ \left[n + 2 \sum\limits_{i',j'\ (i'<j')} r_{i'j'}\right]}} \quad \ldots\ldots\ldots (41)$$

and formula (40) as

$$R = \frac{\sum r_{ii'} + 2 \sum\limits_{i,j\ (i<j)} r_{ij}}{n + 2 \sum\limits_{i,j\ (i<j)} r_{ij}} \quad \ldots\ldots\ldots (42)$$

Formulae (41) and (42) give the reliability of the sum of standard

scores. If the tests in our battery are weighted according to any system of weights, $w_1, w_2, \ldots . w_n$, then the reliability of the sum of weighted scores is given by

$$R = \frac{\sum r_{ii'} w_i^2 + 2 \sum\limits_{i,j'\ (i<j')} r_{ij'} w_i w_j}{\sqrt{\left[\sum w_i^2 + 2 \sum\limits_{i,j\ (i<j)} r_{ij} w_i w_j\right]\ \left[\sum w_i^2 + 2 \sum\limits_{i',j'\ (i'<j')} r_{i'j'} w_i w_j\right]}} \quad \ldots\ldots (43)$$

If we make the equivalence assumption of equation (39), formula (43) may be written as

$$R = \frac{\sum r_{ii'} w_i^2 + 2 \sum\limits_{i,j\ (i<j)} r_{ij} w_i w_j}{\sum w_i^2 + 2 \sum\limits_{i,j\ (i<j)} r_{ij} w_i w_j} \quad \ldots\ldots (44)$$

Formulae (40), (42) and (44) may be written in such a form that the battery reliability may be obtained without calculating all the correlations between tests. We require, however, the variance of the sum of scores, the test variances, and the test reliabilities. The variance of the sum of raw scores may be written

$$S_R^2 = \sum S_i^2 + 2 \sum_{i,j\ (i<j)} r_{ij} S_i S_j \quad \ldots\ldots (45)$$

where S_R^2 is the variance of the sum of raw scores. Hence formula (40) may be written

$$R = \frac{S_R^2 - \sum S_i^2 + \sum r_{ii'} S_i^2}{S_R^2} \quad \ldots\ldots (46)$$

The variance of the sum of standard scores is given by

$$S_S^2 = n + 2 \sum_{i,j\ (i<j)} r_{ij} \quad \ldots\ldots (47)$$

Hence formula (42) becomes

$$R = \frac{S_S^2 - n + \sum r_{ii'}}{S_S^2} \quad \ldots\ldots (48)$$

The variance of the sum of weighted scores may be written

$$S_w^2 = \sum w_i^2 + 2 \sum_{i,j\ (i<j)} r_{ij} w_i w_j \quad \ldots\ldots (49)$$

We may, therefore, write formula (44) as follows:

$$R = \frac{S_w^2 - \sum w_i^2 + \sum r_{ii'} w_i^2}{S_w^2} \quad \ldots\ldots (50)$$

The error variance of the sum of raw scores, assuming equations (38) and (39), is given by

$$S_{e_R}^2 = \sum S_i^2 - \sum r_{ii} S_i^2 \quad \ldots\ldots (51)$$

The error variance of the sum of standard scores is given by

$$S_{e_S}^2 = n - \sum r_{ii} \qquad \ldots\ldots\ldots (52)$$

and the error variance of scores weighted by any given system of weights by

$$S_{e_w}^2 = \sum w_i^2 - \sum r_{ii} w_i^2 \qquad \ldots\ldots\ldots (53)$$

where $S_{e_R}^2$, $S_{e_S}^2$, and $S_{e_w}^2$ denote the error variances of raw, standard, and weighted scores respectively.

For computational purposes in applying the formulae given above it is most convenient to write all the intercorrelations in the form of a pooling square as described by Thomson [111, 112], and described briefly elsewhere in this Bulletin (see p. 72).

The Split-half Reliability of a Test Battery

The split-half reliability of a test battery may be calculated by computing the correlation between the odd and even items of each test, boosting this coefficient by the Spearman-Brown formula, and, when raw scores are used, applying formula (40). The assumption is made that the variances of the odd and even items of any given test do not differ significantly. This assumption is, of course, implicit in the Spearman-Brown formula. If the variances of the halves of the same test differ somewhat the following formula may be used:

$$R = \frac{S_R^2 - \sum^{2n} S_{\frac{i}{2}}^2 + \sum^{2n} r_{\frac{ii}{22}} S_{\frac{i}{2}}^2}{S_R^2} \qquad \ldots\ldots\ldots (54)$$

In the above formula S_R^2 is obtained either by applying formula (45) or by straightforward calculation. The term $\sum^{2n} S_{\frac{i}{2}}^2$ is the sum of the variances of the half tests. The summation is, of course, over $2n$ values. The term $\sum^{2n} r_{\frac{ii}{22}} S_{\frac{i}{2}}^2$ is the sum of the self covariances of test halves. There are n values of $r_{\frac{ii}{22}}$, and $2n$ values of $S_{\frac{i}{2}}^2$, hence each value of $r_{\frac{ii}{22}}$ appears twice in the summation.

Maximum Battery Reliability

Regression weights may be assigned to any given battery of tests to obtain maximum prediction of an external criterion, that is, given a dependent variate z_0, which is presumed to measure a specified

attribute, and n independent variates, $z_1, z_2, \ldots . z_n$, which measure characteristics of that attribute, weights obtained by the method of least squares may be assigned to the independent variates which will maximize the correlation between the scores on the dependent variate z_0 and the sum of weighted scores on the independent variates. When, however, there are a number of dependent variates weights may be assigned to both dependent and independent variates to maximize the correlation between the sum of weighted scores on the criteria and the sum of weighted scores on the predictors. Hotelling [49] furnished a least square solution to this problem.

Thomson [111] applied Hotelling's solution to the special case of determining weights that would yield maximum battery reliability. That is, if $x_1, x_2, \ldots . x_n$ are scores in standard measure obtained on the first application of a test and $x_{1'}, x_{2'}, \ldots . x_{n'}$ are corresponding scores obtained by giving the test a second time or by the administration of parallel forms, we may write two linear functions

$$L_1 = k_1x_1 + k_2x_2 + \quad . \quad . \quad . \quad + k_nx_n$$
$$L_{1'} = k_{1'}x_{1'} + k_{2'}x_{2'} + \quad . \quad . \quad . \quad + k_{n'}x_{n'}$$

and obtain a series of weights $k_1, k_2, \ldots ., k_n$ and $k_{1'}, k_{2'}, \ldots ., k_{n'}$, such that the correlation between L_1 and $L_{1'}$ is a maximum. If, however, we make the equivalence assumption, that is, if $r_{ij} = r_{i'j'} = r_{i'j} = r_{ij'}$, and $k_1 = k_{1'}, k_2 = k_{2'}, \ldots ., k_n = k_{n'}$, then the solution of the required weights is in some degree simplified.

None the less the attainment of weights that will maximize the reliability of a test battery is a matter of much arithmetical labour. If we write, for example, the intercorrelations between two applications of three tests in the form of a matrix, thus,

	z_1	z_2	z_3	$z_{1'}$	$z_{2'}$	$z_{3'}$
z_1	1	r_{12}	r_{13}	r_{11}	r_{12}	r_{13}
z_2	r_{12}	1	r_{23}	r_{12}	r_{22}	r_{23}
z_3	r_{13}	r_{23}	1	r_{13}	r_{23}	r_{33}
$z_{1'}$	r_{11}	r_{12}	r_{13}	1	r_{12}	r_{13}
$z_{2'}$	r_{12}	r_{22}	r_{23}	r_{12}	1	r_{23}
$z_{3'}$	r_{13}	r_{23}	r_{33}	r_{13}	r_{23}	1

and denote the four quadrants by

A	C
C	A

then to attain maximum reliability we must solve the equation

$$|CA^{-1}C-\lambda_1 A|=0 \qquad \ldots\ldots\ldots(55)$$

where λ_1 is the largest latent root. Then

$$R^2=\lambda_1 \qquad \ldots\ldots\ldots(56)$$

where R is the maximum reliability coefficient. The computation of λ_1 with three tests involves the solution of a cubic equation and with n tests the solution of an equation of the nth degree. The weights $k_1, k_2, \ldots, k_n$ which will yield maximum battery reliability are proportional to the elements in any row of the adjugate of the matrix $(CA^{-1}C-\lambda_1 A)$. The above exact solution is a matter of much difficulty when more than three or four tests are included in the calculation.

We may consider, however, a special case of the above general problem. When all tests in a battery are presumed to measure the same attribute, let us say g, and differ only with respect to the accuracy with which this attribute g is measured, then the best estimate of a person's g is given by the linear function

$$\hat{g}=k_1x_1+k_2x_2+\ldots\ldots\ldots+k_nx_n$$

where $k_1, k_2, \ldots, k_n$ are regression weights, and $x_1, x_2, \ldots, x_n$ are in standard measure.

But

$$k_1=\frac{\Delta_{01}}{\Delta_{00}};\ k_2=\frac{-\Delta_{02}}{\Delta_{00}} \qquad \ldots\ldots\ldots(57)$$

where

$$\Delta=\begin{vmatrix} 1 & r_{1g} & r_{2g} & \cdot & \cdot & \cdot & r_{ng} \\ r_{1g} & 1 & r_{12} & \cdot & \cdot & \cdot & r_{1n} \\ r_{2g} & r_{12} & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{ng} & r_{1n} & \cdot & \cdot & \cdot & \cdot & 1 \end{vmatrix}$$

and

$$r_{ij}=r_{ig}r_{jg} \qquad \ldots\ldots\ldots(58)$$

Evaluating Δ_{01} (see [61] pp. 212-213) in terms of r_{1g} we obtain

$$\Delta_{01}=\frac{r_{1g}}{1-r_{1g}}[(1-r_{1g}^2)(1-r_{2g}^2)\ldots.(1-r_{ng}^2)] \qquad \ldots\ldots\ldots(59)$$

But the quantity

$$[(1-r_{1g}^2)(1-r_{2g}^2)\ldots(1-r_{ng}^2)]/\Delta_{00}=\mu$$

where μ is constant for all variates. Hence the relative weights to be

assigned to each test to give maximum prediction of g are given by

$$k_i = \frac{r_{ig}}{1-r_{ig}^2} \qquad \ldots\ldots\ldots (60)$$

But if all our tests are presumed to be a measure of the same attribute, and differ only in the accuracy with which that attribute is measured, that is, if the matrix of intercorrelations may be explained in terms of one general factor and n error specifics, we may write

$$r_{ig}^2 = r_{ii} \qquad \ldots\ldots\ldots (61)$$

hence

$$k_i = \frac{\sqrt{r_{ii}}}{1-r_{ii}} \qquad \ldots\ldots\ldots (62)$$

If we are reasonably satisfied that no specific other than error specific exists in the factorial configuration describing the matrix of intercorrelations, that is, if the equality

$$\frac{r_{ij}}{\sqrt{r_{ii}r_{jj}}} = 1 \qquad \ldots\ldots\ldots (63)$$

is satisfied, we may assign to each of our variates the weights calculated by formula (62) and obtain a best estimate of a man's true score, which under the conditions specified is identical with a best estimate of a man's g. Furthermore these weights are directly proportional to the elements of any row of the adjugate of the matrix $(CA^{-1}C-\lambda_1 A)$, and yield, therefore, maximum battery reliability.

Consider a numerical example. Let the intercorrelations between three tests be as follows:

	z_1	z_2	z_3
z_1	1	.72	.63
z_2	.72	1	.56
z_3	.63	.56	1

and let the reliabilities of the three tests be .81, .64, and .49 respectively. Here the weights computed by formulae (60) and (62) are identical, and are in the ratio

1 .469 .290

The weighted battery has a reliability of .875. The weights obtained by solving the equation

$$|CA^{-1}C-\lambda_1 A| = 0$$

for λ_1 are in corresponding ratio, the value of λ_1 being .7657, and the maximum reliability .875.

If the condition imposed by equation [63] is not satisfied the

three methods yield three different sets of weights, and different battery reliabilities are obtained. Weighting scores by formulae (60) or (62) will usually, although not always, yield a battery reliability greater than the reliability obtained by taking the straight sum of raw or standard scores. In general, we may state that as $r_{ij}^2 \to r_{ii}\, r_{jj}$ or as $r_{ig}^2 \to r_{ii}$ the battery reliability obtained by weighting according to either formula (60) or formula (62) tends towards a maximum.

In the numerical example given by Thomson [111] the table of intercorrelations between three tests was

	z_1	z_2	z_3
z_1	1	.482	.617
z_2	.482	1	.397
z_3	.617	.397	1

the reliabilities of the three tests being .86, .73, and .83 respectively. The weights required to obtain a maximum reliability of .915 are in the ratio

1 .36 .76

The weights obtained by the formula (62) are in the ratio

1 .478 .809

and the obtained battery reliability is .914. The weights that yield a best estimate of g are of the order

1 .234 .420

and the battery reliability resulting from the use of these weights is .910. When equal weights are assigned to each variate the battery reliability is .903. In the above numerical example weighting the variates to obtain maximum reliability does not increase the accuracy of measurement in any great degree.

Reliability and Factor Patterns

The reliability of a test may be written as a function of the factorial structure of the tests included in the battery. Consider firstly the case where the matrix of intercorrelations is of rank 1. If the scores on the n tests included in the battery are in standard measure the battery reliability, R, is given by the formula

$$R = \frac{\sum r_{ii'} + 2\sum_{i,j\,(i<j)} r_{ij}}{n + 2\sum_{i,j\,(i<j)} r_{ij}} \qquad \ldots\ldots\ldots (64)$$

where R = the reliability of the sum of standard scores,

$r_{ii'}$ = the reliability of test i,
r_{ij} = correlation between test i and test j.

If the matrix is of rank 1 the tests in the battery may be described in terms of one general factor, n specifics, and n error specifics.

Writing
$$r_{ij} = a_i a_j$$
$$r_{ii'} = 1-e_i^2$$
$$a_i^2 = 1-S_i^2-e_i^2$$

and
a_i = loading of the ith test in the general factor,
e_i^2 = error variance of the ith test,
S_i^2 = specific variance of the ith test,

then (64) may be written

$$R = \frac{n-\sum_{i=1}^{n} e_i^2+\sum_{i=1}^{n}\sum_{j=1}^{(n-1)}{}_{i\neq j} a_i a_j}{n+\sum_{i=1}^{n}\sum_{j=1}^{(n-1)}{}_{i\neq j} a_i a_j} \qquad \ldots\ldots\ldots(65)$$

which may be put in the form

$$R = \frac{n-\sum_{i=1}^{n} e_i^2+\sum_{i=1}^{n}\sum_{j=1}^{(n-1)}{}_{i\neq j}\sqrt{(1-S_i^2-e_i^2)(1-S_j^2-e_j^2)}}{n+\sum_{i=1}^{n}\sum_{j=1}^{(n-1)}{}_{i\neq j}\sqrt{(1-S_i^2-e_i^2)(1-S_j^2-e_j^2)}} \qquad \ldots\ldots(66)$$

Formula (66) indicates, when the matrix of intercorrelations is of rank 1, the dependency of the battery reliability on the error and specific factor variances.

Consider the case when the rank of the matrix is greater than 1. Let a_i be the loading of test i in the first factor common to a tests, β_i the loading of test i in the second factor common to b tests, and so on. Hence

$$R = \frac{n-\sum e_i^2+\sum_{i=1}^{a}\sum_{j=1}^{(a-1)}{}_{i\neq j} a_i a_j+\sum_{i=1}^{b}\sum_{j=1}^{(b-1)}{}_{i\neq j}\beta_i\beta_j+\ldots\ldots+\sum_{i=1}^{r}\sum_{j=1}^{(r-1)}{}_{i\neq j}\rho_i\rho_j}{n+\sum_{i=1}^{a}\sum_{j=1}^{(a-1)}{}_{i\neq j} a_i a_j+\sum_{i=1}^{b}\sum_{j=1}^{(b-1)}{}_{i\neq j}\beta_i\beta_j+\ldots\ldots+\sum_{i=1}^{r}\sum_{j=1}^{(r-1)}{}_{i\neq j}\rho_i\rho_j} \qquad \ldots\ldots(67)$$

This formula shows the relationship between the reliability of a test battery and the factorial composition of the tests included in it. Obviously from the point of view of reliability no special advantage need be attached to matrices of correlations of rank 1. The reliability of the battery depends in part on the magnitude of the intercorrelations, and although the rank of the matrix is 1, these intercorrelations

may be small. We may observe, however, that as a battery of tests approximates to the measurement of a unit trait the rank of the matrix of intercorrelations tends to unity. The converse does not hold.

Combinatorial Reliability Analysis

Many tests in general use are constructed of sub-tests, each sub-test being more or less homogeneous with respect to content, although the content of different sub-tests may differ substantially. Usually scores on different sub-tests are added together without weighting into a composite score, which is regarded as a function of the ability which the test as a whole presumes to measure.

The reliability of tests of this type, and the consequent stability of the relative order in which persons tested are arranged, depends in part on the reliability of sub-tests, and in part on the relationship between sub-tests. Thus the lengthening of a test by the addition of a new sub-test may not infrequently reduce rather than increase the reliability of measurement.

In designing a test which will arrange persons tested in a reliable order it is relevant to enquire whether or not sub-tests have been included which are exerting a negative influence on the reliability, and to determine which particular combination of sub-tests of all possible combinations yields the most reliable scores. Thus the n sub-tests of a given test of reliability r_{nn} may be taken one at a time and in all combinations to yield n^2-1 possible reliability coefficients. Among these n^2-1 tests and teams of tests may be found a test or team of k tests of reliability r_{kk} such that $r_{kk} \geq r_{ii}$, and $r_{kk} \geq r_{nn}$, where $k \geq i$ and $k \geq n$.

The process by means of which the most reliable combination of k sub-tests in a complex of n sub-tests is determined involves the calculation of the reliability of all possible combinations of sub-tests, and is termed combinatorial reliability analysis.

The process of combinatorial reliability analysis is analogous to the technique of *complete tilling* described by Ragnar Frisch [37] in connection with multiple regression problems. Frisch points out that in enquiries involving multiple regression the prediction attained by the weighted sum of n independent variates is frequently not significantly greater than the prediction attained by the weighted sum of a much smaller number of variates. Since the addition or deletion of independent variates can alter substantially the relative magnitude of the weights to be assigned to the remaining variates, it is necessary

to calculate regression weights for all possible combinations of the independent variates before our information regarding the available data is complete. In multiple regression the addition of new variates can never reduce the multiple correlation coefficient. If, however, the scores are added together without the use of weights, the correlation of the team of independent variates with the criterion may be reduced by the addition of a new variate to the team. Similarly in test reliability the addition of new tests to a battery can never reduce the reliability of the battery weighted to yield maximum reliability. If, however, the scores are added together without the use of weights, as is the common practice, the addition of new tests may reduce significantly the reliability of the battery.

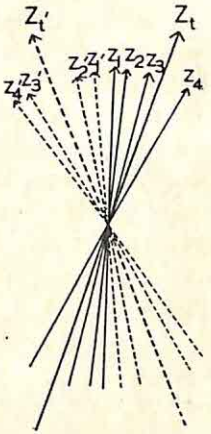

Figure 3.—Geometrical Representation of Battery Reliability.

The reliability of a test consisting of a number of sub-tests may be visualized geometrically as follows. The correlation between two sub-tests may be represented as the cosine of the angle between two vectors, and the intercorrelations between n sub-tests by the cosines of the $(n-1)/2$ angles between a sheaf of n vectors. If our test has been given a second time to the same group of persons, we have two sheaves of vectors as shown in Figure 3 where the unbroken vectors represent the first and the broken vectors the second administration of the group of sub-tests. In Figure 3, for purposes of illustration, our test is presumed to consist of four sub-tests. It will be remembered that there are as many dimensions as there are vectors. The reliability coefficient is the cosine of the angle between the two centroid vectors represented by z_t and $z_{t'}$. By deleting the pair of vectors z_1 and $z_{1'}$ from our

diagram the two centroid vectors are pulled farther apart and the test reliability reduced. By removing vectors z_4 and $z_{4'}$ from our diagram the centroid vectors are drawn closer together, and the reliability of the test increased. The problem visualized geometrically is, therefore, to determine the particular combination of sub-tests that will bring the two centroid vectors as close together as possible.

The reliabilities of all possible combinations of sub-tests may be readily calculated from the matrix of covariances. The covariances by two successive applications of n sub-tests to the same sample of persons may be written in the form of a pooling square, as follows:

	z_1	z_2	. . . . .	z_n	$z_{1'}$	$z_{2'}$	. . . . .	$z_{n'}$
z_1	S_1^2	$r_{12}S_1S_2$	. . .	$r_{1n}S_1S_n$	$r_{11'}S_1S_{1'}$	$r_{12'}S_1S_{2'}$	. . .	$r_{1n'}S_1S_{n'}$
z_2	$r_{12}S_1S_2$	S_2^2	. . . . .		$r_{12'}S_1S_{2'}$	$r_{22'}S_2S_{2'}$	. . . . .	
.	. . . . .				. . . . .			
.	. . . . .				. . . . .			
z_n	$r_{1n}S_1S_n$	. . . . .		S_n^2	$r_{1n'}S_1S_{n'}$	. . . . .		$r_{nn'}S_nS_{n'}$
$z_{1'}$	$r_{1'2}S_{1'}S_1$	$r_{1'2}S_{1'}S_2$	. . .	$r_{1'n}S_{1'}S_n$	$S_{1'}^2$	$r_{1'2'}S_{1'}S_{2'}$	.	$r_{1'n'}S_{1'}S_{n'}$
$z_{2'}$	$r_{1'2}S_{1'}S_2$	$r_{2'2}S_{2'}S_2$	. . . . .		$r_{1'2'}S_{1'}S_{2'}$	$S_{2'}^2$	. . . . .	
.	. . . . .				. . . . .			
.	. . . . .				. . . . .			
$z_{n'}$	$r_{1'n}S_{1'}S_n$	. . . . .		$r_{n'n}S_{n'}S_n$	$r_{1'n'}S_{1'}S_{n'}$	. . . . .		$S_{n'}^2$

The correlation between $z_{(1+2\cdots n)}$ and $z_{(1'+2'\cdots n')}$ is given by dividing the sum of all elements in the North-East quadrant of the above square by the square root of the product of the sum of the elements in the North-West and South-East quadrants, as follows:

$$r_{(1+2\cdots n)\,(1'+2'\cdots n')} = \frac{\sum r_{ii'}S_iS_{i'} + 2\sum\limits_{i,j'(i<j')} r_{ij'}S_iS_{j'}}{\sqrt{[\sum S_i^2 + 2\sum r_{ij}S_iS_j]\,[\sum S_{i'}^2 + 2\sum r_{i'j'}S_{i'}S_{j'}]}} \quad \ldots\ldots\ldots (68)$$

The term in the numerator of the above equation is the covariance; the two terms in the denominator are the variances of the sums of raw scores.

The reliability of all possible combinations of sub-tests may be obtained by calculating values of $\sum r_{ii'}S_iS_{i'} + 2\sum\limits_{i,j'(i<j')} r_{ij'}S_iS_{j'}$, $\sum S_i^2 + 2\sum r_{ij}S_iS_j$ and $\sum S_{i'}^2 + 2\sum r_{i'j'}S_{i'}S_{j'}$ for all combinations of sub-tests, and substituting the values obtained in formula (68).

The method outlined above is illustrated with reference to the following set of data. The Revised Beta Examination was administered twice to three classes of Grade IX pupils ($N=95$). This test is a revision of the United States Army Group Examination Beta. It consists of six sub-tests, described as follows:

Sub-test	Content	Time (min.)	Maximum Score
1	Maze	$1\frac{1}{2}$	10
2	Digit Symbol	2	30
3	Picture Discrimination	3	20
4	Form Board	4	18
5	Picture Completion	$2\frac{1}{2}$	20
6	Number Checking	2	25

All the 66 different intercorrelations between the scores on two successive applications of the six sub-tests were calculated. These intercorrelations, which are given in Table XXII, are observed to be low, a few indeed being negative, indicating low internal consistency. The matrix of covariances with variances in the principal diagonal is given in Table XXIII. From this covariance matrix all statistics necessary to carry out a complete combinatorial analysis may be computed. The variances, covariances, and reliabilities of all possible combinations of sub-tests are given in columns 2, 3, 4, and 5 of Tables XXIVA, XXIVB and XXIVC.

To illustrate how these figures are calculated consider the combination of sub-tests 124. The variance of the scores on the first application of these three sub-tests is obtained by adding together the elements in the N.W. quadrant of the covariance matrix of Table XXIII after deleting rows and columns 3, 5, and 6. This variance is found to be 26.227. The variance of the scores on the second application is calculated in similar manner from the S.E. quadrant, and is 21.685. The covariance 18.385 is obtained in similar manner from the N.W. quadrant. Hence the reliability of the 124 combination, .771, is calculated by dividing the covariances thus obtained by the square root of the product of the two variances.

TABLE XXII

Matrix of Correlations
Revised Beta Examination

	z_1	z_2	z_3	z_4	z_5	z_6	$z_{1'}$	$z_{2'}$	$z_{3'}$	$z_{4'}$	$z_{5'}$	$z_{6'}$
z_1	——	.2477	.2487	.0735	.2194	.0787	.5847	.2029	.0793	.1456	.2696	.0357
z_2	.2477	——	−.0095	−.0347	.0835	.3852	.2539	.7307	−.1596	.0092	.0921	.4813
z_3	.2487	−.0095	——	.2391	.4504	.0355	.1858	−.0421	.6496	.3432	.4563	.0283
z_4	.0735	−.0347	.2391	——	.1365	.0546	.1153	−.0953	.2356	.7953	.1776	−.0699
z_5	.2194	.0835	.4504	.1365	——	.0867	.2522	.0931	.3927	.1434	.8226	.0703
z_6	.0787	.3852	.0355	.0546	.0867	——	.0607	.4642	.0666	.1572	.0503	.6737
$z_{1'}$	.5847	.2539	.1858	.1153	.2522	.0607	——	.0976	.1056	.2244	.3058	.1898
$z_{2'}$	.2029	.7307	−.0421	−.0953	.0931	.4642	.0976	——	−.1221	−.0483	−.0535	.5328
$z_{3'}$	.0793	−.1596	.6496	.2356	.3927	.0666	.1056	−.1221	——	.2907	.3403	.0100
$z_{4'}$	.1456	.0092	.3432	.7953	.1434	.1572	.2244	−.0483	.2907	——	.2574	.1120
$z_{5'}$	.2696	.0921	.4563	.1776	.8226	.0503	.3058	−.0535	.3403	.2574	——	.1275
$z_{6'}$	.0357	.4813	.0283	−.0699	.0703	.6737	.1898	.5328	.0100	.1120	.1275	——

TABLE XXIII

MATRIX OF COVARIANCES WITH VARIANCES IN THE DIAGONAL
REVISED BETA EXAMINATION

	z_1	z_2	z_3	z_4	z_5	z_6	$z_{1'}$	$z_{2'}$	$z_{3'}$	$z_{4'}$	$z_{5'}$	$z_{6'}$
z_1	**1·8235**	1.1856	.8309	.3078	.7072	.2526	.6933	.8572	.2535	.6019	.7734	.1112
z_2	1.1856	**12·5682**	−.0833	−.3816	.7051	3.2467	1.0032	8.1045	−1.3391	.1039	.6939	3.9354
z_3	.8309	−.0833	**6·1240**	1.8341	2.6537	.2087	.5027	−.3259	3.8058	2.5996	2.3995	.1614
z_4	.3078	−.3816	1.8341	**9·6112**	.9205	.4024	.3984	−.9241	1.7293	7.5466	1.1699	−.5000
z_5	.7072	.7051	2.6537	.9205	**5·6686**	.4907	.6693	.6938	.3927	1.0448	4.1612	.3860
z_6	.2526	3.2467	.2087	.4024	.4907	**5·6523**	.5084	3.4525	.3746	1.1440	.2540	3.6936
$z_{1'}$	.6933	1.0032	.5027	.3984	.6693	.5084	**1·2426**	.3402	.2786	.7657	.7243	.4880
$z_{2'}$	.8572	8.1045	−.3259	−.9241	.6938	3.4525	.3402	**9·7878**	−.9040	−.4622	−.1778	3.8438
$z_{3'}$	.2535	−1.3391	3.8058	1.7293	.3927	.3746	.2786	−.9040	**5·6044**	2.1066	1.7119	.5458
$z_{4'}$	.6019	.1039	2.5996	7.5466	1.0448	1.1440	.7657	−.4622	2.1066	**9·3676**	1.6467	.7907
$z_{5'}$	.7734	.6939	2.3995	1.1699	4.1612	.2540	.7243	−.1778	1.7119	1.6467	**4·5144**	.6247
$z_{6'}$	.1112	3.9354	.1614	−.5000	.3860	3.6936	.4880	3.8438	.5458	.7907	.6247	**5·3183**

TABLE XXIVa

DATA OBTAINED FROM COMBINATORIAL RELIABILITY ANALYSIS, REVISED BETA EXAMINATION

Combination of Variables	S_1^2	$S_{1'}^2$	$r_{11'}S_1S_{1'}$	$r_{11'}$
1	1.824	1.243	.693	.585
2	12.568	9.788	8.105	.731
3	6.124	5.604	3.806	.650
4	9.611	9.368	7.547	.795
5	5.669	4.514	4.161	.823
6	5.652	5.318	3.694	.674
12	16.763	11.711	10.658	.761
13	9.609	7.404	5.255	.623
14	12.050	12.142	9.240	.764
15	8.907	7.206	6.297	.786
16	7.981	7.537	5.007	.645
23	18.526	13.584	10.245	.646
24	21.416	18.231	14.831	.751
25	19.647	13.947	13.653	.825
26	24.714	22.794	19.185	.808
34	19.403	19.185	15.681	.813
35	17.100	13.543	10.759	.707
36	12.194	12.014	8.035	.664
45	17.121	17.175	13.923	.812
46	16.068	16.267	11.884	.735
56	12.302	11.082	8.495	.728

TABLE XXIVB

DATA OBTAINED FROM COMBINATORIAL RELIABILITY ANALYSIS, REVISED BETA EXAMINATION

Combination of Variables	S_1^2	$S_{1'}^2$	$r_{11'}S_1S_{1'}$	$r_{11'}$
123	24.382	16.064	13.555	.685
124	26.227	21.685	18.385	.771
134	23.504	22.516	18.131	.788
234	31.042	26.241	21.301	.746
125	25.256	17.318	17.650	.879
135	22.000	16.791	13.651	.710
235	30.912	21.167	18.586	.727
145	20.974	21.398	17.059	.805
245	30.336	25.683	22.595	.810
345	32.220	30.416	24.849	.794
126	29.414	25.693	22.359	.813
136	16.184	14.790	10.105	.653
236	31.089	27.682	21.863	.745
146	19.013	20.017	14.197	.728
246	34.367	32.818	26.556	.791
346	26.278	27.177	20.555	.769
156	16.045	14.749	11.250	.731
256	32.774	28.202	25.375	.835
356	24.151	21.202	15.629	.691
456	24.559	25.325	18.900	.758

TABLE XXIVc

DATA OBTAINED FROM COMBINATORIAL RELIABILITY ANALYSIS, REVISED BETA EXAMINATION

Combination of Variables	S_1^2	$S_{1'}^2$	$r_{11'}S_1S_{1'}$	$r_{11'}$
1234	37.514	30.252	25.611	.760
1235	38.183	25.096	23.339	.794
1245	36.561	30.583	27.591	.825
1345	37.736	35.197	28.742	.789
2345	45.269	37.114	31.856	.777
1236	37.450	31.138	25.792	.772
1246	39.682	37.249	30.730	.799
1346	30.884	31.484	23.624	.758
2346	44.408	41.920	33.562	.778
1256	32.395	32.550	29.991	.924
1356	29.556	25.426	19.141	.698
2356	44.456	35.264	30.844	.779
1456	34.170	29.898	22.656	.709
2456	44.268	41.520	34.960	.816
3456	40.076	39.657	30.363	.762
12345	53.156	42.577	37.609	.791
12346	51.387	46.907	38.492	.784
12356	52.232	41.419	36.216	.779
12456	50.998	47.396	40.576	.825
13456	46.097	45.413	34.875	.762
23456	59.618	54.045	44.758	.789
123456	68.010	60.481	51.130	.797

Examination of the reliability coefficients given in Tables XXIVA, XXIVB, and XXIVC indicates that for heterogeneous material the addition of one or more sub-tests to an existing sub-test or team of sub-tests may increase, decrease, or leave unaltered the reliability. Consider the following sequence:

Sub-tests	r_{11}
5	.823
25	.825
125	.879
1256	.924

Sub-test 5 is the most reliable of all single sub-tests, and is indeed more reliable than the whole test. The reliability of the whole test is .797. Thus we have a situation where a test requiring only $2\frac{1}{2}$ minutes working time, and containing 20 items is more reliable than a combination of 6 sub-tests requiring altogether 15 minutes working time, and containing 123 items. Thus the efficacy of measurement is impaired by some characteristic of the interaction of the tests in combination. Consider the following sequence:

Sub-tests	r_{11}
6	.674
56	.728
156	.731
1356	.698

The reliability of the sequence 1356 is .698 as compared with .924 for the sequence 1256. Thus the addition of test 3 to the 156 combination reduces the reliability coefficient from .731 to .698, while the addition of test 2 to the 156 combination increases the reliability to .924. With heterogeneous material of the type contained in this test no very substantial general tendency exists for the reliability to increase with increase in the number of sub-tests. Indeed the addition of new sub-tests is found in many cases to result in reduced reliability.

The test material used in obtaining the above illustrative data was pictorial in type, and the various sub-tests differed widely in content with the result that the intercorrelations between sub-tests were low. When, however, the intercorrelations between sub-tests are high we should expect to find in general a marked increase in the reliability coefficients obtained by combinatorial analysis with increase in the number of sub-tests.

To illustrate this latter situation forms A and B of the Junior Dominion Group Test of Intelligence were administered to 107 children in Grades IV and V. This test is verbal in type and consists of 5 sub-tests.

Sub-test	Content	No. of items	Time (min.)
1	Opposites	17	3
2	Classification	17	3
3	Analogies	17	3
4	Arithmetic Reasoning	15	7
5	Following Directions	15	9

The items of forms A and B of this test were paired for difficulty and discriminatory power. The two forms are regarded as attaining a high degree of equivalence.

All 45 intercorrelations between the scores obtained on the five sub-tests of these two forms were obtained. These intercorrelations, which are given in Table XXV are seen to be roughly of the order .5 and .6. Table XXVI gives the matrix of covariances. From this matrix as previously described all statistics necessary to compute the reliability of all possible combinations of sub-tests may be readily obtained. The variances, covariances, and reliabilities of all sub-tests and possible combinations of sub-tests are given in columns 2, 3, 4, and 5, respectively, of Tables XXVIIA and XXVIIB.

Here we find a strong tendency for the reliability coefficients to increase with increase in the number of sub-tests. The highest coefficient obtained is for the 1235 combination, .9479, a coefficient which is slightly higher than the coefficient .9211 obtained for the whole test. Apparently the addition of sub-test 4, arithmetical reasoning, exerts a slightly negative effect.

TABLE XXV

Matrix of Correlations
Junior Dominion Group Test of Intelligence

	z_1	z_2	z_3	z_4	z_5	$z_{1'}$	$z_{2'}$	$z_{3'}$	$z_{4'}$	$z_{5'}$
z_1	——	.5170	.4432	.5033	.5062	.6649	.6114	.5394	.5042	.5676
z_2	.5170	——	.6271	.5476	.5606	.5787	.7706	.6335	.5628	.5909
z_3	.4432	.6271	——	.4255	.4181	.4391	.6364	.7245	.4289	.4556
z_4	.5033	.5476	.4255	——	.5661	.5126	.6148	.5243	.8443	.5539
z_5	.5062	.5606	.4181	.5661	——	.4935	.5940	.5029	.5494	.7558
$z_{1'}$	.6649	.5787	.4391	.5126	.4935	——	.5629	.5493	.5662	.5455
$z_{2'}$	.6114	.7706	.6364	.6148	.5940	.5629	——	.6886	.6151	.5723
$z_{3'}$	.5394	.6335	.7245	.5243	.5029	.5493	.6886	——	.4908	.5782
$z_{4'}$	.5042	.5628	.4289	.8443	.5494	.5662	.6151	.4908	——	.5443
$z_{5'}$	.5676	.5909	.4556	.5539	.7558	.5455	.5723	.5782	.5443	——

TABLE XXVI

MATRIX OF COVARIANCES WITH VARIANCES IN THE DIAGONAL
JUNIOR DOMINION GROUP TEST OF INTELLIGENCE

	z_1	z_2	z_3	z_4	z_5	$z_{1'}$	$z_{2'}$	$z_{3'}$	$z_{4'}$	$z_{5'}$
z_1	**8.1735**	5.1281	5.1964	4.4610	4.4878	5.7658	5.9354	5.7720	5.3851	5.2512
z_2	5.1281	**12.0362**	8.9253	5.8911	6.0324	6.0908	9.5169	8.2272	7.2962	6.6360
z_3	5.1964	8.9253	**16.8228**	5.4109	5.3182	5.4642	9.2922	11.1242	6.5732	6.0488
z_4	4.4610	5.8911	5.4109	**9.6131**	5.4436	4.8211	6.7851	6.0845	9.7139	5.5587
z_5	4.4878	6.0324	5.3182	5.4436	**9.6207**	4.6426	6.5572	5.6325	6.3669	7.5870
$z_{1'}$	5.7658	6.0908	5.4642	4.8211	4.6426	**9.2035**	6.0785	6.2377	6.4175	5.3563
$z_{2'}$	5.9354	9.5169	9.2922	6.7851	6.5572	6.0785	**12.6686**	9.1751	8.1820	6.5930
$z_{3'}$	5.7720	8.2272	11.1242	6.0845	5.6325	6.2377	9.1751	**14.0141**	6.8644	7.0050
$z_{4'}$	5.3851	7.2962	6.5732	9.7139	6.3669	6.4175	8.1820	6.8644	**13.9603**	6.5820
$z_{5'}$	5.2512	6.6360	6.0488	5.5587	7.5870	5.3563	6.5930	7.0050	6.5820	**10.4757**

TABLE XXVIIA

DATA OBTAINED FROM COMBINATORIAL RELIABILITY ANALYSIS, JUNIOR DOMINION GROUP TEST OF INTELLIGENCE

Combination of Variables	S_1^2	$S_{1'}^2$	$r_{11'}S_1S_{1'}$	$r_{11'}$
1	8.174	9.204	5.766	.665
2	12.036	12.669	9.517	.771
3	16.823	14.014	11.124	.725
4	9.613	13.960	9.714	.844
5	9.621	10.476	7.587	.756
12	30.466	34.029	27.309	.875
13	35.389	35.693	28.126	.791
14	26.709	35.999	25.686	.875
15	26.770	30.392	23.247	.815
23	46.710	45.033	38.161	.832
24	33.432	42.993	33.312	.879
25	33.722	36.331	30.297	.866
34	37.258	41.703	33.496	.850
35	37.080	38.500	30.393	.804
45	30.121	37.600	29.227	.869

TABLE XXVIIB

DATA OBTAINED FROM COMBINATORIAL RELIABILITY ANALYSIS, JUNIOR DOMINION GROUP TEST OF INTELLIGENCE

Combination of Variables	S_1^2	$S_{1'}^2$	$r_{11'}S_1S_{1'}$	$r_{11'}$
123	75.532	78.869	67.189	.871
124	60.783	77.188	61.310	.895
134	64.746	76.217	60.704	.864
234	78.927	89.086	74.613	.890
125	61.127	68.404	57.983	.897
135	64.622	70.891	57.288	.846
235	79.032	82.705	70.622	.874
145	56.192	70.351	55.092	.876
245	66.004	79.819	66.018	.910
345	68.402	79.353	64.690	.878
1234	116.671	135.757	113.848	.905
1235	116.830	127.253	115.573	.948
1245	102.332	124.727	103.910	.920
1345	104.866	124.579	101.792	.891
2345	122.136	139.922	119.001	.910
12345	168.856	197.305	168.129	.921

Whenever the scores on sub-tests, or the scores on tests in a battery, are added together without the use of weights the technique of combinatorial reliability analysis may be employed to determine whether the test is functioning efficiently. Obviously if one or two sub-tests are more reliable than all sub-tests combined the scores should not be combined by simple additive methods.

If the test has not been given a second time, *consistency coefficients* calculated either by Hoyt's method of analysis or by the Kuder and Richardson formula (20) may be used in carrying out the combinatorial analysis. Under such circumstances coefficients are obtained which indicate the properties of the various parts of the test to coexist with one another.

The calculation required for a complete combinatorial analysis using consistency coefficients is somewhat simpler than the process of calculation already described in this chapter. If consistency coefficients are used the elements in the upper left-hand quadrant of the covariance matrix are identical with the elements in the lower right-hand quadrant, and, with the exception of the diagonal elements, identical with the elements in the upper right-hand quadrant.

The error variance of a single test is given by

$$S_{e_i}^2 = S_i^2(1-r_{ii})$$

The variances for any combination of tests are directly additive. Hence the consistency coefficient of any k tests is given by

$$r_{tt} = 1 - \frac{\sum^{k} S_i^2(1-r_{ii})}{2\sum_{i,j\ (i<j)}^{k} r_{ij} S_i S_j} \qquad \cdots\cdots\cdots (69)$$

Thus consistency coefficients for all possible combinations of tests may be readily calculated.

CHAPTER VII

THE REPORTING OF DATA RELATING TO THE RELIABILITY OF A TEST

Complete and detailed information concerning the reliability of a test is a major factor not only in enabling the research worker to determine a test's usefulness in dealing with a particular problem, but also to assist him to interpret the results obtained. The acquisition of reasonably complete data regarding a test's reliability requires the planning and execution of one or more special experiments, the collection and analysis of a considerable body of data, and the presentation, preferably in the Manual of Directions accompanying the test, of the findings of such analysis. The test maker will see, therefore, that the assessment of a test's reliability, and consequent efficiency as a measuring instrument, is a matter for specific and detailed investigation.

In the previous chapters we have discussed various problems relating to reliability, and their solutions. The main findings will be summarized here, however, in order that the reader may have the necessary details clearly in mind when considering our suggestions regarding the reporting of data dealing with the reliability of a test. Not all the suggestions refer directly to the problem of reliability. Clearly a certain amount of descriptive material must be given, such as the type of test, material used, purpose of the test, etc., and a discussion of the type of problems in which the test may be used. In addition it must be clearly stated in which situations the test is to be used, for which grade or other unit, and the information relating to reliability must be given separately for each separate unit in order that the reader may determine the usefulness of the test in any given situation.

Any report on the reliability of a test must give an estimate of both the absolute and the relative accuracy with which the test measures. To determine the absolute accuracy we need a knowledge of the magnitude of the errors of measurement; this can be given most conveniently in the form of what is called the standard error of measurement. To determine the relative accuracy we require a knowledge of the relative magnitude of these errors of measurement in comparison with the magnitude of the differences between the individuals (or groups) tested. This can be given in the form of either the usual reliability coefficient or the sensitivity coefficient of the test; the latter form is

preferred by the authors. Other information relating to the distribution of the scores, etc., should also be given.

In reporting on the reliability of a test a clear distinction must be drawn between results which refer to the reliability in the usual sense and the internal consistency of the test. In the latter case the test is given once and a function of the scores used to give an estimate of the internal consistency of the test; this is not necessarily the same as reliability. In the case of reliability either the same test is given twice, or two comparable forms of the test are given to the same individuals at different times and some function of the scores of the same individuals on the two trials used as a measure of reliability. It is suggested that measures of both internal consistency and reliability should be given, and any discrepancy between the results explained.

A choice between two statistical methods of analysing the experimental results will frequently have to be made in both of the cases considered in the previous paragraph. In the past the correlation technique has been used almost exclusively but in many cases, as has been shown in an earlier chapter, the analysis of variance and covariance method is to be preferred. There is no incompatibility between the results of the two analyses, and in most cases the choice of which of the two methods to employ will be determined by the nature of the problem under consideration. In most of the problems connected with the reliability and internal consistency of a test, the authors prefer the analysis of variance and covariance method for the following reasons:

1. The arithmetical operations involved in the analysis appear to be simpler and easier to carry out.
2. Considerably more information is extracted from the data and made available for use when required, and the necessary tests of significance are made with less difficulty.
3. The results are easier to interpret and hence are less liable to lead to the drawing of unwarranted conclusions.

For these reasons, therefore, the authors suggest that the analysis of variance and covariance method should be used in analysing the results unless the nature of the problem is such that the use of the correlation technique is indicated, e.g. in the combinatorial reliability analysis. The analysis of variance and covariance method is, of course, a general one and is applicable to any or all of these problems.

There is, finally, the question of the analysis appropriate for special kinds of tests. For the omnibus type of test the usual kind of analysis

can be employed but in battery tests, on the other hand, a series of analyses should be given. For a battery test we need to know the reliability and internal consistency of each sub-test, the relationship between the various sub-tests, and, in addition, the reliability and internal consistency of all possible combinations of sub-tests. For this type of test, therefore, it is suggested that a combinatorial reliability analysis should be reported and also, wherever practicable, a study made of the question of weighting the sub-tests to give maximum reliability of the battery.

Below we have prepared an outline of a report on information which in our opinion should be presented in the Manual of Directions accompanying a test. If any reader feels that a report of the type suggested will involve a large amount of unnecessary work, or if he is of the opinion that the details to be reported are more of academic than of practical interest, he is asked to carry out a simple enquiry. Let him select a problem in some educational field involving the use of tests of types which have already been constructed, and let him then endeavour to determine from the available information which tests are applicable in solving his particular problem, and with what degree of accuracy he may expect to measure under the conditions specified by his problem. In the majority of cases his experience will no doubt be similar to ours. Our ordinary work involves the use of many tests, and also the furnishing of advice to teachers and others regarding the most suitable test or tests to use in the solution of a particular problem. Rarely have we been able to answer all relevant questions by reference to the published information. Usually we find it necessary either to plan and carry out a special experiment of our own, or recommend for use a test or tests by a particular author whose work we know to be in general satisfactory. We suggest, therefore, that the information outlined in the following summary be collected and published by the author of each test. The reporting of a single reliability coefficient without even a statement of the number of persons tested or a description of the group tested contributes little or nothing to our knowledge.

No mention has been made in the report given below of reliability coefficients calculated by the split-half or the odd-even method. We suggest that this method of estimating the reliability or the internal consistency of tests be no longer employed. Other methods are available which do not involve the arbitrary division of the test into parts, and yield substantially more information.

SUGGESTED TEST REPORT (IN MANUAL OF DIRECTIONS)

A. *General Statement*

1. *Purpose of test:*

2. *Situations in which it is recommended for use:*
   e.g. Grade or grades, ages, etc., of pupils.

3. *Time taken to administer:*
   Working time for test proper and total time.

4. *Type of test:*
   Battery or omnibus,
   Self-administering or otherwise,
   Method of pupil-response.

5. *Number of items:*
   By sub-tests and total,
   Method of scoring,
   Possible scores on sub-tests and total.

6. *Test material:*
   Description of type of material used in each sub-test.

7. *Norms:*
   Type of norms given and how these are to be interpreted,
   Description of population to which norms apply,
   Description of method used in calculating norms.

B. *Data Relating to Study of Reliability*

1. *Number of pupils tested:*
   By age, sex, grade and other units considered.

2. *Distribution of scores:*
   Give actual distribution of scores (plus mean and standard deviation) for each grade or other unit considered, and for each sub-test and the whole test.

3. *Reliability:*

Separately for each grade or sampling unit considered.

(*a*) *Internal consistency*

Suggest use of method proposed by Hoyt [51].

(*b*) *Test-retest*

Suggest use of method proposed by Jackson [52].

(*c*) *Comparable forms*

Suggest use of method proposed by Jackson [52].

(*Note:*—In addition the usual correlation coefficients may be given although they add nothing to the information obtained by the use of the above methods.)

4. *Battery tests:*

In the case of a battery of tests or sub-tests, all the information under 3 above is to be given for each sub-test. In addition, all the intercorrelations of the sub-tests and a combinatorial reliability analysis are to be reported.

5. *Standard error of an individual score:*

To be reported for each sub-test and total test and for each grade or other unit employed.

6. *Intelligence tests:*

Report the above information separately for raw scores, mental age scores, and intelligence quotients if these different kinds of scores yield different results.

7. *Comparison with other tests:*

Report correlation coefficients and other relevant data.

APPENDICES

APPENDIX A

NOTE ON THE ESTIMATION OF RELIABILITY COEFFICIENTS

In order that the results given in this Bulletin might be directly comparable with those given elsewhere, the general formula for the Pearson product-moment correlation coefficient has been used in calculating the reliability coefficients. As was mentioned earlier, however, the use of this particular formula is not always justified in problems such as these. The sample value is used as an estimate of the population value, and hence the particular method to be used in calculating the value of our estimate will be determined by the conditions which are assumed to exist in the population from which we are sampling. If the conditions are changed, then obviously we must make an appropriate change in the method used in estimating the population values. This section is devoted to a discussion of these problems and to the development of the methods to be used in the estimation of reliability coefficients under different conditions.

It is necessary, in the first place, to point out that underlying the correlation method is the fundamental assumption that the variables which we are correlating are normally distributed and that the regressions are linear in form. It follows that when we use this method we assume, implicitly or explicitly, that these conditions are satisfied. Fortunately, in much of the educational work of this kind, these conditions are roughly satisfied. Our tests are so constructed that, if they are used in the appropriate situations, few individuals make very low or very high scores and the general distribution may be adequately represented by a normal curve. Similarly, the linearity of regression condition is generally satisfied. If the test is too easy or too difficult for the pupils, however, these conditions are not satisfied, but it is assumed that research workers will be careful to use tests only in the situations for which they were designed. If the conditions are not satisfied, then it is generally impossible to interpret our statistical results but this limitation is, of course, well-known.

In the discussion which follows, we shall assume that the variables are normally distributed and the regressions are linear. It must not be concluded, however, that because of this the results deduced here are less valid than the ones more widely used. They apply, in fact, to all situations in which correlation methods may be used.

Let us consider the general case in which we have two sets of scores on the same test for the same individuals. As far as the general discussion is concerned, it is immaterial whether they refer to results of the test-retest, alternative forms or split-half experimental method. If we assume that in the population from which we are sampling the variables are normally distributed, then the form of the distribution will, in general, be determined by the values of the two means, the two standard deviations and the correlation coefficient. This is the general case to which the usual Pearson product-moment formula for estimating the correlation coefficient applies, and it will be considered first.

There are, however, two other cases in which, it is maintained, the conditions agree more closely with those which exist in the problem of estimating the reliability coefficients. The second case to be considered is one in which the two standard deviations may be assumed to be equal: in this case the form of the distribution of the variables in the population from which we are sampling will be determined by the values of the two means, the common standard deviation and the correlation coefficient. These conditions are assumed to hold in the test-retest and alternative forms experimental methods, and in analysing these experimental results we must use the appropriate estimates of the population values. The third case to be considered is the one in which it may be assumed that the two means are equal, and the two standard deviations are equal: in this case the form of the distribution of the variables in the population from which we are sampling will be determined by the values of the common mean, the common standard deviation and the correlation coefficient. These conditions are assumed to hold in the split-half experimental method and, again, in analysing these experimental results we must use the appropriate estimates of the population values.

In the following discussion, each of these cases will be considered separately. The "maximum likelihood" method will be used in each case to determine the best estimates of the population values. This well-known method is based on the simple assumption that we should use as estimates those values which maximize the probability of the observed event; hence the name, "maximum likelihood" method. To obtain the estimates we differentiate the probability function with respect to the parameters in which we are interested, set the resulting equations equal to zero, and solve. This process gives us what are termed the maximum likelihood estimates of the parameters or population values.

Case 1. *General Case*

Denote by X_i and Y_i the scores obtained by the i-th individual on the first and second trials of the test, respectively; by M_x and M_y the means, by σ_x and σ_y the standard deviations, and by ρ the correlation coefficient in the sampled population of X and Y. The subscripts, x or y, denote the variables to which these parameters refer. If the variables are normally distributed, then if we denote by $p(X_i)$, $p(Y_i)$ and $p(X_i, Y_i)$ the probability distributions of X_i, Y_i, and X_i and Y_i, respectively, it is known that

$$p(X_i) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(X_i-M_x)^2}{2\sigma_x^2}} \qquad \ldots\ldots\ldots(70)$$

$$p(Y_i) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(Y_i-M_y)^2}{2\sigma_y^2}} \qquad \ldots\ldots\ldots(71)$$

$$p(X_i, Y_i) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{(X_i-M_x)^2}{\sigma_x^2} - 2\rho\frac{(X_i-M_x)(Y_i-M_y)}{\sigma_x\sigma_y} + \frac{(Y_i-M_y)^2}{\sigma_y^2}\right\}} \qquad \ldots\ldots\ldots(72)$$

If we have N pairs of values, then the simultaneous probability distribution of all the N values of X_i and Y_i will be

$$p(X_1, \ldots, X_N, Y_1, \ldots, Y_N) =$$

$$= \left(\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\right)^N e^{-\frac{1}{2(1-\rho^2)}\Sigma\left\{\frac{(X_i-M_x)^2}{\sigma_x^2} - 2\rho\frac{(X_i-M_x)(Y_i-M_y)}{\sigma_x\sigma_y} + \frac{(Y_i-M_y)^2}{\sigma_y^2}\right\}} \quad \ldots\ldots\ldots(73)$$

where Σ denotes summation for all N values of i.

As we wish to estimate M_x, M_y, σ_x, σ_y and ρ, we shall have to take the partial derivatives of $p(X_1, \ldots, X_N, Y_1, \ldots, Y_N)$, which we may denote by p for short, with respect to each of these. It is more convenient, however, to work with the natural logarithm of p: this will not affect the results, of course, as p will be a maximum if $\log p$ is a maximum. We have

$$\log p = -N\log 2\pi - N\log\sigma_x - N\log\sigma_y - \frac{N}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\Sigma\left\{\frac{(X_i-M_x)^2}{\sigma_x^2} - \frac{2\rho(X_i-M_x)(Y_i-M_y)}{\sigma_x\sigma_y} + \frac{(Y_i-M_y)^2}{\sigma_y^2}\right\} \quad \ldots\ldots\ldots(74)$$

Differentiating $\log p$ with respect to M_x, we obtain:

$$\frac{\delta\log p}{\delta M_x} = 0 = \frac{2}{2(1-\rho^2)}\frac{\Sigma(X_i-M_x)}{\sigma_x^2} - \frac{2\rho}{2(1-\rho^2)}\frac{\Sigma(Y_i-M_y)}{\sigma_x\sigma_y} \quad \ldots\ldots\ldots(75)$$

which reduces to

$$\sigma_y\Sigma(X_i-M_x) = \rho\sigma_x\Sigma(Y_i-M_y) \quad \ldots\ldots\ldots(76)$$

Similarly, differentiating $\log p$ with respect to M_y, setting the equation equal to zero and simplifying, we obtain:

$$\sigma_x\Sigma(Y_i-M_y) = \rho\sigma_y\Sigma(X_i-M_x) \quad \ldots\ldots\ldots(77)$$

Assuming $\sigma_x \neq 0$, $\sigma_y \neq 0$, $\rho \neq 1$, we may solve (76) and (77) to obtain

$$\left.\begin{aligned} M_x &= \frac{\Sigma X_i}{N} \\ M_y &= \frac{\Sigma Y_i}{N} \end{aligned}\right\} \quad \ldots\ldots\ldots(78)$$

Differentiating $\log p$ partially with respect to σ_x, σ_y, ρ in turn, setting the resulting equations equal to zero, solving and substituting the values given in (78) for M_x and M_y, we have finally:

$$\left.\begin{aligned} \sigma_x &= \sqrt{\frac{1}{N}\left\{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N}\right\}} \\ \sigma_y &= \sqrt{\frac{1}{N}\left\{\Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{N}\right\}} \end{aligned}\right\} \quad \ldots\ldots\ldots(79)$$

$$\rho = \frac{\Sigma X_iY_i - \frac{(\Sigma X_i)(\Sigma Y_i)}{N}}{\sqrt{\left\{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N}\right\}\left\{\Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{N}\right\}}} \quad \ldots\ldots\ldots(80)$$

The maximum likelihood estimates of the five parameters, M_x, M_y, σ_x, σ_y, and ρ, are given, therefore, in equations (78), (79) and (80). These require no explanation as they are, of course, the estimates generally used.

Case 2. *Equal Standard Deviations*

In this case, we assume that the standard deviations of the two distributions are equal, i.e.

$$\sigma_x = \sigma_y = \sigma \qquad \ldots\ldots\ldots (81)$$

where σ denotes the value of the standard deviation common to the two distributions. The simultaneous probability distribution, p, of all the N values of X_i and Y_i will be

$$p = \left(\frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}}\right)^N e^{-\frac{1}{2\sigma^2(1-\rho^2)}\Sigma\left\{(X_i-M_x)^2-2\rho(X_i-M_x)(Y_i-M_y)+(Y_i-M_y)^2\right\}} \qquad \ldots\ldots\ldots (82)$$

Using log p; differentiating with respect to M_x, M_y, σ, ρ, separately; setting the resulting equations equal to zero, and solving as in the previous case, we find:

$$\left.\begin{aligned} M_x &= \frac{\Sigma X_i}{N} \\ M_y &= \frac{\Sigma Y_i}{N} \\ \sigma &= \sqrt{\frac{1}{2N}\left\{\left(\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N}\right) + \left(\Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{N}\right)\right\}} \\ \rho &= \frac{2\left\{\Sigma X_i Y_i - \frac{(\Sigma X_i)(\Sigma Y_i)}{N}\right\}}{\left(\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N}\right) + \left(\Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{N}\right)} \end{aligned}\right\} \qquad \ldots\ldots\ldots (83)$$

The values given in equation (83) are the maximum likelihood estimates of the four parameters M_x, M_y, σ and ρ. It follows, therefore, that in problems in which the assumption of a common standard deviation is satisfied, these are the appropriate estimates of the population values. In analysing the data obtained by using the test-retest and alternative forms experimental methods, we should, therefore, use the formulae given in (83) in calculating the estimates of the parameters in which we are interested.

Case 3. *Equal Means and Equal Standard Deviations*

In this case we assume that the means and the standard deviations of the two distributions are equal, i.e.

$$\left.\begin{aligned} M_x &= M_y = M \\ \sigma_x &= \sigma_y = \sigma \end{aligned}\right\} \qquad \ldots\ldots\ldots (84)$$

where M denotes the value of the mean, and σ the value of the standard deviation, common to the two distributions. Proceeding as before, we obtain finally:

$$\left.\begin{aligned} M &= \frac{\Sigma X_i + \Sigma Y_i}{N} \\ \sigma &= \sqrt{\frac{1}{2N}\left\{\Sigma X_i^2 + \Sigma Y_i^2 - \frac{(\Sigma X_i + \Sigma Y_i)^2}{2N}\right\}} \\ \rho &= \frac{2\Sigma X_i Y_i - \frac{(\Sigma X_i + \Sigma Y_i)^2}{2N}}{\Sigma X_i^2 + \Sigma Y_i^2 - \frac{(\Sigma X_i + \Sigma Y_i)^2}{2N}} \end{aligned}\right\} \quad \ldots\ldots\ldots (85)$$

The values given in equation (85) are the maximum likelihood estimates of the three parameters M, σ, ρ. In problems in which we may assume a common mean and standard deviation for the two distributions, we should, therefore, use these formulae in calculating the estimates of the parameters. It follows, therefore, that in analysing the data obtained by the use of the split-half experimental method, we should use the formulae given in equation (85) in calculating the estimates of the population values.

Examples

(1) The values given below refer to the scores made by 29 pupils on two forms of an intelligence test. To save space, only the necessary totals are given:

$$\begin{aligned} N &= 29 \\ \Sigma X_i &= 633 \\ \Sigma Y_i &= 757 \\ \Sigma X_i^2 &= 16537 \\ \Sigma Y_i^2 &= 23685 \\ \Sigma X_i Y_i &= 19269 \end{aligned}$$

To test whether or not the standard deviations may be assumed to be equal, we calculate

$$F = \frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N}}{\Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{N}} = 1.44$$

and refer to Snedecor's tables of F with degrees of freedom $n_1 = n_2 = 28$. We find that F is less than the 5% point given in the table of F, so we conclude that the assumption of a common standard deviation is justified.

Using the formulae given in equation (83), we find that the best estimate of the reliability coefficient is $r = 0.826$. If we had used the formula given in equation (80), the formula usually used, we should have had $r = 0.840$ as our estimate of the reliability coefficient. The difference between the estimates is not large, and in some cases is even smaller, but in cases of this kind we must, to be consistent, use the best available estimate of the reliability coefficient.

(2) The values given below refer to the scores made by 29 pupils on the odd and even items of a test. The totals are:

$$\begin{aligned} N &= 29 \\ \Sigma X_i &= 327 \\ \Sigma Y_i &= 316 \\ \Sigma X_i^2 &= 3731 \\ \Sigma Y_i^2 &= 3478 \\ \Sigma X_i Y_i &= 3591 \end{aligned}$$

We find, making the appropriate tests, that neither the means nor the standard deviations of the two distributions are significantly different; we may; therefore, assume a common mean and a common standard deviation.

Using the formula given in equation (85), we find $r=0.665$ as the estimate of the reliability coefficient of half the test. If we use the formula given in equation (80), we find $r=0.714$ as the estimate of the reliability coefficient of half the test. The difference here is considerable, and clearly in cases such as these we must be careful to use the appropriate formula in calculating our estimates.

It should be pointed out that in all cases we must test whether or not our assumptions of equal standard deviations, or means, are justified. Sometimes they are not, and in these cases it is necessary to find the reason. In using the split-half method, for example, we found an interesting case in which the mean score on the odd items was not equal to the mean score on the even items. When we calculated the difficulties of the items, we found that they were not arranged in order of difficulty; many of the odd items were considerably harder than the corresponding even items. Obviously, for such a test one cannot use the split-half method in determining the reliability.

APPENDIX B

NOTE ON TESTS OF CERTAIN HYPOTHESES RELATING TO THE PROBLEM OF MEASURING THE SENSITIVITY OF A MENTAL TEST

In a previous paper, Jackson [52] has considered the problem of measuring the reliability or consistency of mental tests and suggested a measure, termed the sensitivity of the test, based on the concept of the relative accuracy of the measurements. The present section is concerned with the development of tests of certain hypotheses closely related to this problem.

Underlying the solution of the above problem is the assumption that the score of an individual on the s-th trial of a test may be considered as the sum of certain factors or components, i.e

$$Y_{st} = A + B_s + C_t + z_{st} \qquad \ldots\ldots\ldots (86)$$

where $s = 1, 2$; $t = 1, 2, \ldots . n$; n represents the number of pupils and Y_{st} the score of the t-th individual on the s-th trial of the test. A is considered as a measure of the common ability of the group. It is assumed to be constant, and is defined as the arithmetic mean of the "true" effects for all trials and individuals. B_s is considered as a measure of the trial effect, and is also assumed to be a constant. C_t is considered as a measure of the individual effect, i.e. a measure of some capacity or ability of the t-th individual. z_{st} is a measure of the residual or error effect, i.e. a measure of the errors of measuring by means of the test. It is assumed, further, that in the population from which we are sampling, C_t is normally distributed about zero with constant standard deviation, σ_c, constant for all trials, and z_{st} is normally distributed about zero with constant standard deviation, σ, constant for all trials and individuals.

The sensitivity of a test, denoted by γ, is defined as the ratio of these two standard deviations, i.e

$$\gamma = \frac{\sigma_c}{\sigma} \qquad \ldots\ldots\ldots (87)$$

Since γ expresses the differences between the individuals tested in terms of the errors of measurement of the test, it may be called a measure of the relative accuracy of the test. For this reason, the interpretation of the sensitivity is particularly simple and easy to understand.

The problems discussed in the above paper are as follows:

(1) to determine if there is a significant trial effect;
(2) to determine if the mental test actually measures the capacity of the individuals tested;
(3) to estimate the trial effect if it exists;
(4) to estimate the sensitivity of the mental test if it is found that the test measures the capacity of the individuals tested.

The statistical hypotheses to be tested in the solution of problems (1) and (2) are

$$\begin{aligned} H_1 &: B_s = 0 \\ H_2 &: C_t = 0, (\gamma = 0) \end{aligned} \qquad \ldots\ldots\ldots (88)$$

The procedure to be followed in testing these hypotheses, together with the solution for problems (3) and (4), has been given elsewhere in this Bulletin (Chapter III) and will not be repeated here. In using this method, however, it has been found that there is another problem, similar to (2) in many respects, which it is convenient to consider before we proceed to the estimation of the sensitivity of the test. It arises only when we reject $H_2 : C_t = 0$ (or $\gamma = 0$) and therefore conclude that the test does measure the capacity of the individuals. It may happen that we find $\gamma > 0$, i.e. we reject H_2, but that the test is not sensitive enough to give results which could be considered as satisfactory. If $\gamma = 1$, for example, then $\sigma_e = \sigma$ and we would conclude that the error effect might be as important as the individual effect (i.e. the capacity of the individual) in determining the actual score obtained by the individual. We should, in fact, make an error as great or greater than σ_e units in about 32% of cases in using the scores as estimates of the true capacity of the individuals. The problem of selecting a lower limit for the sensitivity of tests, say $\gamma = 2$, for example, in order that tests which do not reach this standard may be automatically eliminated as unsatisfactory, is a psychological rather than a statistical problem. It is clear that the particular value chosen as the lower limit of sensitivity will be determined by the conditions of the experiment and the use which is to be made of the results. A test may give satisfactory results in one situation but comparatively useless results in another. For these reasons the particular numerical value of γ to be used has not been specified in the theoretical part of this paper, but the general case $\gamma = K$, where K is some constant greater than zero, has been considered.

The statistical hypothesis to be tested may be stated in the form

$$H_3 : \gamma = K \qquad \ldots\ldots\ldots (89)$$

where K is some specified constant, always greater than zero. We should, of course, always test the hypothesis $H_2 : \gamma = 0$ first; if we accept H_2 there will be no need to carry the analysis further. The purpose of this section, therefore, is to consider the problem,

(5) to determine if the mental test is *sensitive enough* to yield satisfactory results.

We shall follow the method outlined in Chapter III and use the theory of testing statistical hypotheses devised and developed by Neyman and Pearson [75, 76].

The set of hypotheses alternative to H_3 will include all hypotheses specifying values of γ greater than zero but not equal to K. If we reject the hypothesis $H_3 : \gamma = K$, however, we shall generally wish to know whether the true value of γ is greater than or less than K. If γ appears to be less than K, we would conclude that the test is not sensitive enough to yield satisfactory results; if, on the other hand, γ appears to be greater than K we would conclude that the test is sensitive enough (more sensitive than is necessary, in fact) to yield satisfactory results although in both cases we would reject the hypothesis $H_3 : \gamma = K$. It seems necessary, therefore, to distinguish between the two sets of hypotheses alternative to H_3, namely:

(Case 1) the set of alternative hypotheses specifying values of γ less than K, i.e. $H_3^1 : \gamma = K^1$ where K^1 is some constant greater than zero but less than K.

(Case 2) the set of alternative hypotheses specifying values of γ greater than K, i.e. $H_3^{11}: \gamma = K^{11}$ where K^{11} is some constant greater than K.

It will be necessary, also, to develop two separate tests of the hypothesis $H_3: \gamma = K$, one for each set of alternative hypotheses.

We shall consider the case of two trials ($s = 1, 2$ only) as the data are generally collected in this form. Since A was considered as a constant and defined as the arithmetic mean of the effects for all trials and individuals, it follows that

$$\begin{aligned} \sum_t C_t &= 0 \\ \sum_s B_s &= 0 \qquad \text{.....(90)} \\ \text{or} \quad B_1 &= -B_2 = -B, \text{ say} \end{aligned}$$

where B is now the measure of the trial effect. We may rewrite (86) in the form

$$\left.\begin{aligned} Y_{1t} &= A - B + C_t + z_{1t} \\ Y_{2t} &= A + B + C_t + z_{2t} \end{aligned}\right\} \qquad \text{.....(91)}$$

From the above assumptions it is easily seen that

(*a*) Y_{1t} is normally distributed about a mean $A - B$ with constant standard deviation equal to $\sqrt{\sigma^2 + \sigma_c^2}$ and

(*b*) Y_{2t} is normally distributed about a mean $A + B$ with the same constant standard deviation.

Denoting by $p(Y_{st})$ the elementary probability distribution of Y_{st}, we have

$$p(Y_{1t}) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\sigma_c^2}} e^{-\frac{(Y_{1t}-A+B)^2}{2(\sigma^2+\sigma_c^2)}} \qquad \text{.....(92)}$$

$$p(Y_{2t}) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\sigma_c^2}} e^{-\frac{(Y_{2t}-A-B)^2}{2(\sigma^2+\sigma_c^2)}} \qquad \text{.....(93)}$$

It follows that the simultaneous probability distribution of all the Y's is

$$p(Y_{11}, \ldots Y_{2n}) = \left(\frac{1}{2\pi(\sigma^2+\sigma_c^2)\sqrt{1-\rho^2}}\right)^n exp\left[-\frac{1}{2(1-\rho^2)}\left\{\sum_{t=1}^{n}\left(\frac{(Y_{1t}-A+B)^2}{\sigma^2+\sigma_c^2}\right.\right.\right.$$
$$\left.\left.\left. -2\rho\frac{(Y_{1t}-A+B)(Y_{2t}-A-B)}{\sigma^2+\sigma_c^2} + \frac{(Y_{2t}-A-B)^2}{\sigma^2+\sigma_c^2}\right)\right\}\right] \qquad \text{.....(94)}$$

where ρ is the correlation coefficient in the sampled population of Y's.

Define
$$\theta = 1 + 2\gamma^2 \qquad \text{.....(95)}$$

and consider the simultaneous distribution of the $2n$ other variables defined by the equations:

$$\begin{aligned} u_t &= Y_{1t} + Y_{2t} \\ v_t &= Y_{2t} - Y_{1t} \end{aligned} \qquad \text{.....(96)}$$

Making the appropriate transformation, and substituting

$$\rho = \frac{\gamma^2}{1+\gamma^2} \qquad \text{.....(97)}$$

we find that u_t and v_t are distributed independently and

$$p(u_1, \ldots, u_n) = \left(\frac{1}{\sqrt{2\pi}\sqrt{2\theta}\sigma}\right)^n e^{-\frac{1}{4\sigma^2\theta}\left[\sum_t (u_t-\bar{u})^2 + n(\bar{u}-2A)^2\right]} \quad \ldots\ldots (98)$$

$$p(v_1, \ldots, v_n) = \left(\frac{1}{\sqrt{4\pi}\sigma}\right)^n e^{-\frac{1}{4\sigma^2}\left[\sum_t (v_t-\bar{v})^2 + n(\bar{v}-2B)^2\right]} \quad \ldots\ldots (99)$$

where

$$\left.\begin{aligned} \bar{u} &= \frac{1}{n}\sum_t u_t \\ \bar{v} &= \frac{1}{n}\sum_t v_t \end{aligned}\right\} \quad \ldots\ldots (100)$$

Define, also,

$$\left.\begin{aligned} nS_u^2 &= \sum_t \left\{u_t - \bar{u}\right\}^2 \\ nS_v^2 &= \sum_t \left\{v_t - \bar{v}\right\}^2 \end{aligned}\right\} \quad \ldots\ldots (101)$$

It may be shown that if the value of θ is specified, then the statistics

$$\left.\begin{aligned} T_1 &= \bar{u} \\ T_2 &= \bar{v} \\ T_3 &= \frac{S_u^2}{\theta} + S_v^2 \end{aligned}\right\} \quad \ldots\ldots (102)$$

form a sufficient set for the parameters A, B and σ. The hypothesis to be tested, H_3, specifies the value $\gamma_0 = K$ of γ, or the value $\theta_0 = 1 + 2K^2$ of θ. The two sets of alternative hypotheses, H_3^1 and H_3^{11}, specify the values of θ less than θ_0 and greater than θ_0, respectively.

Case 1. *Test of the Hypothesis H_3: $\gamma = K$ for the set of Alternative Hypotheses H_3^1: $\gamma < K$*

The hypothesis H_3 does not specify the values of three parameters, A, B, and σ, and is, therefore, a composite hypothesis. The critical region for testing H_3 must be "similar" [74] to the sample space with respect to A, B, and σ, that is to say, such that the probability of its containing the sample E be equal to some chosen value ϵ, independent of the values that those parameters may possess. The general method of constructing such regions [77] if a set of sufficient statistics exists, consists of the following.

Denote by W the whole sample space and by the single letter T the whole set of statistics sufficient for A, B, and σ. Further, let $W(T)$ be the locus of points where T is constant and let $w(T)$ denote a part of $W(T)$. Give, in turn, all possible values to the statistics forming the sufficient set T and combine the regions $w(T)$ into one region w. If the region $w(T)$ is chosen with the restriction that the probability that E fall in $w(T)$, calculated on the assumption that it

fall in $W(T)$, be equal to ϵ, so that $P\{E\epsilon w(T)/E\epsilon W(T)\}=\epsilon$, then the region w will be similar to the sample space with respect to A, B, and σ.

If it is desired that this region w be most powerful with respect to some particular alternative hypothesis H_3^1, it is necessary and sufficient that the regions $w(T)$ be chosen on each $W(T)$ to satisfy the inequality:

$$p(Y_{11}, \ldots, Y_{2n}/H_3^1) \geq k(T_1, T_2, T_3)p(Y_{11}, \ldots, Y_{2n}/H_2) \quad \ldots\ldots\ldots(103)$$

where $k(T_1, T_2, T_3)$ is a function of T_1, T_2, and T_3 only.

It is easy to see that on each surface $W(T)$ the inequality (103) reduces to

$$S_u^2 \leq k_1(T_1, T_2, T_3) \quad \ldots\ldots\ldots(104)$$

where $k_1(T_1, T_2, T_3)$ is again a function of T_1, T_2, and T_3 only and is to be determined to satisfy the condition $P\{E\epsilon w(T)/E\epsilon W(T)\}=\epsilon$. Since the assumption that H_3 is true and T_3 has a fixed value implies that S_u^2 cannot exceed θT_3, this last condition reduces to

$$\int_0^{k_1} p(S_u^2, T_1, T_2, T_3)\, dS_u^2 = \epsilon \int_0^{\theta_0 T_3} p(S_u^2, T_1, T_2, T_3)\, dS_u^2 \quad \ldots\ldots\ldots(105)$$

where $\theta_0 = 1+2K^2$ and ϵ denotes the probability of errors of the first kind which we fix beforehand.

The probability distribution $p(S_u^2, T_1, T_2, T_3)$ may easily be obtained from the distribution given in equations (98) and (99). Substituting the value of $p(S_u^2, T_1, T_2, T_3)$ in equation (105) and simplifying, we find

$$\int_0^{k_1} (S_u^2)^{\frac{n-3}{2}} \left(T_3 - \frac{S_u^2}{\theta_0}\right)^{\frac{n-3}{2}} dS_u^2 = \epsilon \int_0^{\theta_0 T_3} (S_u^2)^{\frac{n-3}{2}} \left(T_3 - \frac{S_u^2}{\theta_0}\right)^{\frac{n-3}{2}} dS_u^2 \quad \ldots\ldots\ldots(106)$$

Let
$$S_u^2 = \theta_0 T_3 z \qquad \ldots\ldots\ldots(107)$$
then
$$dS_u^2 = \theta_0 T_3 dz$$

substituting in (106) and simplifying, we obtain

$$\frac{\Gamma_{(n-1)}}{\left(\Gamma_{\frac{n-1}{2}}\right)^2} \int_0^{\frac{k_1}{\theta_0 T_3}} z^{\frac{n-3}{2}} (1-z)^{\frac{n-3}{2}} dz = \epsilon \quad \ldots\ldots\ldots(108)$$

We can find $\frac{k_1}{\theta_0 T_3} = z_0$, say, from the Tables of the Incomplete Beta-Function [80] for any given value of ϵ. We reject the hypothesis to be tested when

$$S_u^2 \leq k_1 = z_0 \theta_0 T_3 \quad \ldots\ldots\ldots(109)$$

But $T_3 = \frac{S_u^2}{\theta_0} + S_v^2$, and substituting this value in equation (109) we find that we reject the hypothesis to be tested when

$$\frac{S_u^2}{S_v^2} \leq \frac{\theta_0 z_0}{1-z_0} \quad \ldots\ldots\ldots(110)$$

Or, alternatively, we may use an adaptation of the z-test of R. A. Fisher; i.e. Calculate the value

$$\left.\begin{aligned} z_3 &= \tfrac{1}{2}\log_e\left\{\frac{S_u^2}{S_v^2}\right\} \\ f_1 &= f_2 = n-1 \end{aligned}\right\} \qquad \ldots\ldots\ldots(111)$$

and reject the hypothesis to be tested, $H_3 : \gamma = K$ (or $\theta = \theta_0$), when z_3 is less than either

$$\left.\begin{aligned} z_{3(5\%)} &= \tfrac{1}{2}\log_e(1+2K^2) - z_{5\%} \\ z_{3(1\%)} &= \tfrac{1}{2}\log_e(1+2K^2) - z_{1\%} \end{aligned}\right\} \qquad \ldots\ldots\ldots(112)$$

where $z_{5\%}$ and $z_{1\%}$ are the 5% and 1% points, respectively, given in Fisher's tables of the distribution of z [32].

As this test does not depend on the alternative value of θ, it is uniformly most powerful with respect to all the alternatives of the class considered.

Case 2. *Test of the Hypothesis* H_3: $\gamma = K$ *for the set of Alternative Hypotheses* H_3^{11}: $\gamma > K$

Proceeding as in Case 1, we find that a common best critical region exists and is defined by

$$p(Y_{11}, \ldots, Y_{2n}/H_3^{11}) \geq Q(T_1, T_2, T_3)\, p(Y_{11}, \ldots, Y_{2n}/H_3) \qquad \ldots\ldots\ldots(113)$$

where H_3^{11} is a composite hypothesis alternative to H_3 and specifying some value of θ greater than θ_0.

On the surface for which the sufficient set of statistics of equation (102) is constant, equation (113) reduces to

$$S_u^2 \geq Q_1(T_1, T_2, T_3) \qquad \ldots\ldots\ldots(114)$$

where Q_1 (T_1, T_2, T_3), or Q_1 for short, is chosen so as to satisfy

$$\int_{Q_1}^{\theta_0 T_3} p(S_u^2, T_1, T_2, T_3)\, dS_u^2 = \epsilon \int_0^{\theta_0 T_3} p(S_u^2, T_1, T_2, T_3)\, dS_u^2 \qquad \ldots\ldots\ldots(115)$$

where $\theta_0 = 1+2K^2$ and ϵ denotes the probability of errors of the first kind which we fix beforehand.

Substituting the value of $p(S_u^2, T_1, T_2, T_3)$ of the previous section in equation (115) and simplifying, we have

$$\int_{Q_1}^{\theta_0 T_3} (S_u^2)^{\frac{n-3}{2}} \left(T_3 - \frac{S_u^2}{\theta_0}\right)^{\frac{n-3}{2}} dS_u^2 = \epsilon \int_0^{\theta_0 T_3} (S_u^2)^{\frac{n-3}{2}} \left(T_3 - \frac{S_u^2}{\theta_0}\right)^{\frac{n-3}{2}} dS_u^2 \qquad \ldots\ldots\ldots(116)$$

making the transformation given in equation (107), we find that equation (116) reduces to

$$\frac{\Gamma(n-1)}{\left(\Gamma\frac{n-1}{2}\right)^2} \int_{\frac{Q_1}{\theta_0 T_3}}^1 z^{\frac{n-3}{2}} (1-z)^{\frac{n-3}{2}}\, dz = \epsilon \qquad \ldots\ldots\ldots(117)$$

We can find $\dfrac{Q_1}{\theta_0 T_3} = l_0$, say, from the Tables of the Incomplete Beta-Function for any value of ϵ. We reject the hypothesis to be tested when

$$S_u^2 \geq Q_1 = l_0 \theta_0 T_3 \qquad \ldots\ldots\ldots(118)$$

But $T_3 = \frac{S_u^2}{\theta_0} + S_v^2$, and substituting this value in (118) we find that we reject the hypothesis to be tested, $H_3: \gamma = K$, when

$$\frac{S_u^2}{S_v^2} \geq \theta_0 \frac{l_0}{1-l_0} \qquad \text{.....(119)}$$

Or, alternatively, we may use an adaptation of the z-test of R. A. Fisher; i.e. calculate the value

$$\left.\begin{aligned} z_3 &= \tfrac{1}{2} \log_e \left\{ \frac{S_u^2}{S_v^2} \right\} \\ f_1 &= f_2 = n-1 \end{aligned}\right\} \qquad \text{.....(120)}$$

and reject the hypothesis to be tested, $H_3: \gamma = K$, when z_3 is greater than either

$$\left.\begin{aligned} z^1_{3(5\%)} &= \tfrac{1}{2} \log_e (1+2K^2) + z_{5\%} \\ z^1_{3(1\%)} &= \tfrac{1}{2} \log_e (1+2K^2) + z_{1\%} \end{aligned}\right\} \qquad \text{.....(121)}$$

where $z_{5\%}$ and $z_{1\%}$ are the 5% and 1% points, respectively, given in Fisher's tables of the distribution of z.

Examples

The following results refer to a test in French Reading for Grade X, prepared by our Department and given to two classes of pupils.

TABLE XXVIII

RESULTS OF ANALYSIS OF SCORES ON FRENCH READING TEST

	1st Class	2nd Class
n	35	39
nS_u^2	5288.2000	3704.8718
nS_v^2	166.0857	553.8462
Estimate of γ	3.93	1.69
$z_3 = \frac{1}{2} \log_e \left\{ \frac{S_u^2}{S_v^2} \right\}$	1.730	0.950

From Fisher's tables of z, we find

(a) for Class 1

$$f_1 = f_2 = 34$$

$$z_{5\%} = 0.281 \qquad z_{1\%} = 0.401$$

(*b*) for Class 2

$$f_1 = f_2 = 38$$

$$z_{5\%} = 0.264 \qquad z_{1\%} = 0.376$$

Let us assume that we have decided the test will yield satisfactory results if its sensitivity is not less than the arbitrary standard of $\gamma_0 = 2$; we know that if the test is as sensitive as this, then in using the scores as estimates of the true ability of the individuals (in French Reading) we shall make an error as great or greater than σ_c units by chance alone in less than 5% of cases. The hypothesis we wish to test, therefore, is $H_3: \gamma = 2$ (i.e. $K = 2$).

Consider, first, the test of the hypothesis $H_2: C_t = 0$, i.e. $\gamma = 0$. It will be seen from the values of z_3 given in Table XXVIII and the values of the 5% and 1% points given above, that we would reject this hypothesis in both cases. We conclude, therefore, that the test does measure the ability of the individuals, and we may proceed to the test of the hypothesis $H_3: \gamma_0 = 2$ (i.e. $K = 2$).

It will be seen from the values given as estimates of γ in Table XXVIII, that for Class 1 we must consider the set of alternative hypotheses specifying values of γ greater than 2, and for Class 2 we must consider the set of alternative hypotheses specifying values of γ less than 2. The tests for the two classes are considered separately in the following analysis.

(*a*) *Class 1. Test of the Hypothesis* $H_3: \gamma_0 = 2$

Since the set of hypotheses alternative to H_3 which we consider specify values of γ greater than $\gamma_0 = 2$, we use the results given above under Case 2. We find

$$z^1_{3(5\%)} = \tfrac{1}{2} \log_e (1+2K^2) + z_{5\%}$$
$$= 1.099 + 0.281 = 1.380$$
$$z^1_{3(1\%)} = \tfrac{1}{2} \log_e (1+2K^2) + z_{1\%}$$
$$= 1.099 + 0.401 = 1.500$$

We find that the observed value of $z_3 = 1.730$ is greater than the 1% point so we reject the hypothesis tested, $H_3: \gamma_0 = 2$. We conclude that the test gives results which are satisfactory since it proves to have a sensitivity which is significantly higher than our selected standard.

(*b*) *Class 2. Test of the Hypothesis* $H_3: \gamma_0 = 2$

Since the set of hypotheses alternative to H_3 which we consider specify values of γ less than $\gamma_0 = 2$, we use the results given above under Case 1. We find

$$z_{3(5\%)} = \tfrac{1}{2} \log_e (1+2K^2) - z_{5\%}$$
$$= 1.099 - 0.264 = 0.835$$
$$z_{3(1\%)} = \tfrac{1}{2} \log_e (1+2K^2) - z_{1\%}$$
$$= 1.099 - 0.376 = 0.723$$

We see that the observed value of $z_3 = 0.950$ is greater than the 5% point so we accept the hypothesis tested, $H_3: \gamma_0 = 2$.

Our conclusion, therefore, should be that there is no evidence against the hypothesis tested, $H_3: \gamma_0 = 2$ and this may be considered as a justification for

applying the test. However, there is a considerable difference between the situation concerning the classes 1 and 2.

It should be noted that the tests of $H_3: \gamma = K$ are not very powerful in the case of small samples. The distributions on which our tests of significance are based overlap to a marked extent for samples in which f_1 and f_2 are less than 100. It follows that the probability of errors of the second kind, i.e. the probability of our accepting the hypothesis tested when it is, in fact, false, will be very large for small samples, if we consider alternative hypotheses which specify values of γ not very different from the value, $\gamma_0 = K$, specified by the hypothesis to be tested. The tests given above, however, are the best possible in the sense that the probability of errors of the first kind is controlled at a fixed level and the probability of errors of the second kind is reduced to as low a level as possible for the type of hypotheses in which we are interested.

All this is relevant from the point of view of interpreting the above numerical results. In both cases of class 1 and class 2 the conclusion is favourable to the test, but because of chance variation it may, in fact, be wrong. If it is wrong for class 1, it would mean that we committed an error of the first kind. But the probability of this, in our case, does not exceed 1 per cent. On the other hand, if the presumption that the test is satisfactory and $\gamma_0 = 2$ is wrong for class 2, then it would mean that in our statistical analysis we committed an error of the second kind. Because of what has been said above about the power of the test, this last circumstance is not so unlikely.

APPENDIX C

NOTE ON THE RELATIONSHIP BETWEEN RELIABILITY COEFFICIENTS CALCULATED FROM MENTAL AGE AND I.Q. SCORES

When we speak of the reliability of a test, say for a particular school grade, we imply that there is one and only one reliability coefficient for a particular test. Ignoring the effect of the variability in the population sampled, there are at least two and in some cases three reliability coefficients which we may calculate and use. Our estimate of the reliability may refer to (1) the raw scores, (2) the mental age scores or (3) the I.Q. scores. Since for most tests there is a simple linear relationship between the raw scores and the mental age equivalents, the first two estimates will generally be the same. This is not always the case, however, as for some tests there is not such a simple relationship between raw scores and mental age scores. This difference is not as important as the difference between the second and third estimates, and hence the present note is concerned only with the relationship between the latter two.

In a recent paper [54], Jackson has discussed the general relationship between mental age and I.Q. scores so there is no need to repeat the arguments here. It was shown that the correlation between I.Q. scores and chronological age will generally be negative, and that the correlation between two sets of mental age scores will not be equal, except under certain special conditions, to the correlation between the corresponding I.Q. scores unless the influence of chronological age is removed (by the usual partial correlation technique) from both coefficients.

Since a reliability coefficient is only a special form of a correlation coefficient, these results may be applied directly to the present problem. When the population from which we are sampling is a school grade, the correlation between the I.Q. scores and chronological age will generally be negative, and very often large. For mental age scores and chronological age, on the other hand, the correlation will generally be positive but low, and in many cases will not be significantly different from zero. It follows, therefore, that the reliability coefficients calculated from mental age and I.Q. scores will not necessarily be the same and we cannot speak simply of *the* reliability of a test.

The reader may immediately raise the question as to which reliability coefficient should be given. As a matter of fact there is no general answer to this question because some workers may wish to use mental age scores and others I.Q. scores. To ensure general satisfaction, all the coefficients should be given in order that a worker may use the value appropriate to his particular problem.

Some examples of the kind of results which may be obtained are given below. The comparable forms estimates of reliability are shown in Table XXIX for each four grades and the total for all grades, and the correlation between mental age, I.Q. scores and chronological age for Grade 5 only in Table XXX. Although the samples are small, these results are typical.

TABLE XXIX

COMPARISON OF RELIABILITY COEFFICIENTS CALCULATED FROM MENTAL AGE AND I.Q. SCORES (BY GRADES)

Grade	Number of Pupils	Reliability Coefficients	
		Mental Age	I.Q.
7	38	0.865	0.953
6	42	0.907	0.974
5	40	0.830	0.937
4	38	0.888	0.961
Total	158	0.934	0.958

TABLE XXX

RELATIONSHIP BETWEEN MENTAL AGE, I.Q. SCORES AND CHRONOLOGICAL AGE (GRADE 5 ONLY, 40 CASES)

Form used	Correlation with Chron. Age	
	Mental Age	I.Q.
Form A	+0.036	−0.766
Form B	−0.056	−0.815

If we correct the Grade 5 reliability coefficients for chronological age, we have a value of 0.834 for mental age and 0.839 for I.Q. scores.

When we sample from a population composed of several grades, the difference between the mental age and I.Q. reliability coefficients tends to disappear. In this case, of course, we increase greatly the variability of both mental and chronological age but the variability of the I.Q. scores is relatively unchanged. The values shown in the last row of Table XXIX illustrate this point; the reliability coefficient based on I.Q. scores is still higher but the difference is not as great as in the other cases.

APPENDIX D

NOTE ON THE RELATIONSHIP BETWEEN RELIABILITY AND SAMPLING WITHOUT REPLACEMENT

In the succeeding discussion certain reliability formulae are developed from probability considerations used in the theory of sampling without replacement.

Consider, firstly, an urn containing N balls of which Np are white and Nq are black. From this urn we may draw a ball n times *without replacement*, and the variance of blacks or of whites in repeated sampling is given by the formula

$$\sigma^2 = npq - \frac{n(n-1)pq}{N-1} \qquad \ldots\ldots\ldots (122)$$

where p = proportion of white balls in population sampled;
q = proportion of black balls in population sampled;
n = number drawn;
N = number of balls in population.

We assume in the above formula that the balls are unbiased.

Now an intelligence test constructed of N items, $N/2$ odd and $N/2$ even, may be likened to an urn containing N balls, one half of which are white and the other half black, the variance of white or black balls in repeated sampling without replacement being given by the formula

$$\sigma^2 = \frac{n}{4} - \frac{n(n-1)}{4(n-1)} \qquad \ldots\ldots\ldots (123)$$

In splitting a test by calculating scores on the odd and even items we may argue that, if the test is split without bias, $p=q=\frac{1}{2}$. Thus if a large number of persons make a score X on a test Z of N items, the variance of scores on the odd items, which here is presumed equal to the variance of scores on the even items, is given by

$$\sigma^2_{Z_0} = \frac{X}{4} - \frac{X(X-1)}{4(n-1)} \qquad \ldots\ldots\ldots (124)$$

Since all the balls in the urn to which we likened our test were presumed to exist without bias and to be equally probable, it is clear that formula (124) makes the assumption that all the items of our test must be of the same difficulty.

Since
$$\sigma^2_{Z_0} = \frac{\Sigma Z_0^2}{K} - \frac{X^2}{4} \qquad \ldots\ldots\ldots (125)$$

we may from formula (124) write

$$\frac{\Sigma Z_0^2}{K} = \frac{X}{4} - \frac{X(X-1)}{4(N-1)} + \frac{X^2}{4} \qquad \ldots\ldots\ldots (126)$$

Summing over all values of X and averaging we obtain

$$\frac{\Sigma Z_0^2}{K} = \frac{M}{4} - \frac{\overline{X^2}}{4(N-1)} + \frac{M}{4(N-1)} + \frac{\overline{X^2}}{4} \qquad \cdots\cdots(127)$$

where M is the mean of X.

Hence $$\sigma_{Z_0}^2 = \frac{M}{4} - \frac{\overline{X^2}}{4(N-1)} + \frac{M}{4(N-1)} + \frac{\overline{X^2}}{4} - \frac{M^2}{4} \qquad \cdots\cdots(128)$$

which, since $$\sigma_t^2 = \overline{X^2} - M^2$$

reduces to $$\sigma_{Z_0}^2 = \frac{1}{4(N-1)}\left[M(N-M) + \sigma_t^2(N-2)\right] \qquad \cdots\cdots(129)$$

This formula gives the variance of scores on either the odd or the even items.

But the reliability of a test on the assumption that the variance of scores on the odd items is equal to the variance of scores on the even items is given by

$$r_{tt} = 2 - \frac{4\sigma_{Z_0}^2}{\sigma_t^2} \qquad \cdots\cdots(130)$$

Substituting equation (129) in equation (130) we obtain

$$r_{tt} = \frac{N}{N-1}\left[1 - \frac{M(N-M)}{N\sigma_t^2}\right] \qquad \cdots\cdots(131)$$

This formula is identical with the Kuder-Richardson formula (21). It is probable that the Kuder-Richardson formula (20) also can be derived by a method similar to that described above by taking into consideration the fact that the items may be biased.

BIBLIOGRAPHY

1. Ackerson, L. "In Disagreement with E. A. Lincoln's Article, 'The Unreliability of Reliability Coefficients'." *Journal of Educational Psychology*, 24, (1933), pp. 233-235.
2. Adams, Henry F. "Validity, Reliability and Objectivity." *Psychological Monographs*, 47, (1936), pp. 329-350.
3. Anastasi, A. "Influence of Practice upon Test Reliability." *Journal of Educational Psychology*, 25, (1934), pp. 321-335.
4. Asker, William. "Reliability of Tests Requiring Alternate Responses." *Journal of Educational Research*, 9, (1924), pp. 234-240.
5. Babitz, Milton, and Keys, Noel. "A Method of Approximating the Average Intercorrelation by Correlating the Parts with the Sum of Parts." *Psychometrika*, 5, (1940), pp. 283-288.
6. Bingham, Walter V. "Reliability, Validity and Dependability." *Journal of Applied Psychology*, 16, (1932), pp. 116-122.
7. Brown, William. "Some Experimental Results in the Correlation of Mental Abilities." *British Journal of Psychology*, 3, (1910), pp. 296-322.
8. Brownell, W. A. "On the Accuracy with which Reliability may be Measured by Correlating Test Halves." *Journal of Experimental Education*, 1, (1933), pp. 204-215.
9. Carr, Harvey A. "The Reliability vs. the Validity of Test Scores." *Psychological Review*, 45, (1938), pp. 435-440.
10. Casanova, Teobaldo. "Analysis of the Effect upon the Reliability Coefficient of Changes in Variables Involved in the Estimation of Test Reliability." *Journal of Experimental Education*, 9, (1941), pp. 219-228.
11. Cleeton, Glen U. "Optimum Difficulty of Group Test Items." *Journal of Applied Psychology*, 10, (1926), pp. 327-340.
12. Copeland, H. A. "Note on the Effect of Teaching on the Reliability Coefficient of an Achievement Test." *Journal of Applied Psychology*, 18, (1934), pp. 711-716.
13. Crum, W. L. "Note on the Reliability of a Test with Special Reference to the Examinations set by the College Entrance Board." *American Mathematical Monthly*, 30, (1923), p. 296.
14. Cureton, Edward E., and Dunlap, Jack W. "Nomograph for Estimating the Reliability of a Test in One Range of Talent when its Reliability is Known in another Range." *Journal of Educational Psychology*, 20, (1929), pp. 537-538.
15. Cureton, Edward E., and Dunlap, Jack W. "A Nomograph for Estimating a Reliability Coefficient by the Spearman-Brown Formula and for Computing its Probable Error." *Journal of Educational Psychology*, 21, (1930), pp. 68-69.
16. Cureton, Edward E. "Errors of Measurement and Correlation." *Archives of Psychology*, 125, (1931), pp. 8-13.
17. Cureton, Edward E. "Validation against a Fallible Criterion." *Journal of Experimental Education*, 1, (1933), pp. 258-263.

18. Cureton, Edward E. "Standard Error of the Spearman-Brown Formula when used to Estimate the Length of a Test Necessary to Achieve a Given Reliability." *Journal of Educational Psychology*, 24, (1933), pp. 305-306.
19. Denney, H. R., and Remmers, H. H. "Reliability of Multiple-Choice Measuring Instruments as a Function of the Spearman-Brown Prophecy Formula, II." *Journal of Educational Psychology*, 31, (1940), pp. 699-704.
20. Dickey, John W. "On the Reliability of a Standard Score." *Journal of Educational Psychology*, 21, (1930), pp. 547-549.
21. Dickey, John W. "On Estimating the Reliability Coefficient." *Journal of Applied Psychology*, 18, (1934), pp. 103-115.
22. Douglass, H. R., and Cozens, F. W. "On Formulae for Estimating the Reliability of Test Batteries." *Journal of Educational Psychology*, 20, (1929), pp. 369-377.
23. Douglass, H. R. "A Note on the Correctness of Certain Error Formulas." *Journal of Educational Psychology*, 20, (1929), pp. 434-437.
24. Douglass, H. R. "A Further Note on the Corrections of Certain Error Formulas." *Journal of Educational Psychology*, 21, (1930), pp. 621-624.
25. Dressel, Paul. "Some Remarks on the Kuder-Richardson Reliability Coefficient." *Psychometrika*, 5, (1940), pp. 305-310.
26. Dunlap, Jack W., De Mello, Adrian, and Cureton, Edward E. "The Effects of Different Directions and Scoring Methods on the Reliability of a True-False Test." *School and Society*, 30, (1929), pp. 378-384.
27. Dunlap, J. W. and Cureton, E. E. "Note on the Standard Error of a Reliability Coefficient for a Different Range of Talent." *Journal of Educational Psychology*, 20, (1929), pp. 705-706.
28. Dunlap, J. W. "Comparable Tests and Reliability." *Journal of Educational Psychology*, 24, (1933), pp. 442-453.
29. Edgerton, Harold A., and Toops, Herbert A. "A Table for Predicting the Validity and Reliability Coefficients of a Test when Lengthened." *Journal of Educational Research*, 18, (1928), pp. 225-234.
30. Ferguson, George A. "A Bifactor Analysis of Reliability Coefficients." *British Journal of Psychology*, 31, (1940), pp. 172-182.
31. Ferguson, George A. *Reliability of Mental Tests.* London: University of London Press, 1941. Pp. 176.
32. Fisher, R. A. *Statistical Methods for Research Workers.* 7th Edition. Edinburgh: Oliver and Boyd. 1938. Pp. xv+356.
33. Flanagan, John C. "Note on Calculating the Standard Error of Measurement and Reliability Coefficients with the Test-Scoring Machine." *Journal of Applied Psychology*, 23, (1937), p. 529.
34. Foran, T. G. "A Note on Methods of Measuring Reliability." *Journal of Educational Psychology*, 22, (1931), pp. 383-387.
35. Franzen, R. H., and Derryberry, M. "Reliability of Group Distinctions." *Journal of Educational Psychology*, 23, (1932), pp. 586-593.
36. Franzen, R. H., and Derryberry, M. "Note on Reliability Coefficients." *Journal of Educational Psychology*, 23, (1932), pp. 559-560.
37. Frisch, Ragner. "Statistical Confluence Analysis by Means of Complete Regression Systems." *Nordic Statistical Journal*, 5, (1934).
38. Garrett, Henry E. *Statistics in Psychology and Education.* 2nd Edition. New York: Longmans, Green & Co., 1937. Pp. xiv+493.

39. Gates, G. S. "Individual Differences as Affected by Practice." *Archives of Psychology*, 58, (1922).
40. Goodenough, F. L. "Critical Note on the Use of the Term Reliability in Mental Measurement." *Journal of Educational Psychology*, 27, (1936), pp. 173-178.
41. Gordon, Kate. "Group Judgements in the Field of Lifted Weights." *Journal of Experimental Psychology*, (1924).
42. Gulliksen, Harold. "The Content Reliability of a Test." *Psychometrika*, 1, (1936), pp. 189-194.
43. Gundlach, R. "The Effect of Practice on the Correlation of Three Mental Traits." *Journal of Educational Psychology*, 17, (1926), pp. 387-401.
44. Handy, Uvan, and Lentz, Theodore F. "Item Value and Test Reliability." *Journal of Educational Psychology*, 25, (1934), pp. 703-708.
45. Holzinger, K. J. "Note on the Use of Spearman's Prophecy Formula for Reliability." *Journal of Educational Psychology*, 14, (1923), pp. 302-305.
46. Holzinger, Karl J., and Clayton, Blythe. "Further Experiments in the Application of Spearman's Prophecy Formula." *Journal of Educational Psychology*, 16, (1925), pp. 289-299.
47. Holzinger, Karl J. "A Note on the Correctness of Certain Error Formulas." *Journal of Educational Psychology*, 20, (1929), pp. 669-670.
48. Holzinger, K. J. "Reliability of a Single Test Item." *Journal of Educational Psychology*, 23, (1932), pp. 411-417.
49. Hotelling, H. "The Most Predictable Criterion." *Journal of Educational Psychology*, 26, (1935), pp. 139-142.
50. Hotelling, H. "Relations Between Two Sets of Variates." *Biometrika*, 28, (1936), pp. 328-377.
51. Hoyt, Cyril. "Test Reliability Obtained by Analysis of Variance." *Psychometrika*, 6, (1941), pp. 153-160.
52. Jackson, Robert W. B. "Reliability of Mental Tests." *British Journal of Psychology*, 29, (1939), pp. 267-287.
53. Jackson, Robert W. B. *Application of the Analysis of Variance and Covariance Method to Educational Problems.* Bulletin No. 11, Department of Educational Research. Toronto: University of Toronto. 1940. Pp. 103.
54. Jackson, Robert W. B. "Some Pitfalls in the Statistical Analysis of Data Expressed in the Form of I.Q. Scores." *Journal of Educational Psychology*, 31, (1940), pp. 677-685.
55. Jones, Edward S. "Reliability in Marking Examinations." *Journal of Higher Education*, 8, (1938), pp. 436-439.
56. Jordan, R. C. "Empirical Study of the Reliability Coefficient." *Journal of Educational Psychology*, 26, (1935), pp. 416-426.
57. Kelley, T. L. "The Reliability of Test Scores." *Journal of Educational Research*, 3, (1921), pp. 370-379.
58. Kelley, T. L. "A New Method for Determining the Difference in Intelligence and Achievement Scores." *Journal of Educational Psychology*, 14, (1923), pp. 321-333.
59. Kelley, T. L. "Note on the Reliability of a Test." *Journal of Educational Psychology*, 15, (1924), pp. 193-204.
60. Kelley, T. L. "The Applicability of the Spearman-Brown Formula for the Measurement of Reliability." *Journal of Educational Psychology*, 16, (1925), pp. 300-303.

61. Kelley, T. L. *Interpretation of Educational Measurements.* New York: World Book Company. 1927. Pp. xiii+363.
62. Kreezer, George L., and Bradway, Katherine P. "The Direct Determination of the Probable Error of Measurement of Binet Mental Age." *Journal of Educational Research*, 33, (1939), pp. 197-214.
63. Kuder, G. F., and Richardson, M. W. "The Theory of the Estimation of Test Reliability." *Psychometrika*, 2, (1937), pp. 151-160.
64. Lanier, Lyle H. "Prediction of the Reliability of Mental Tests and Tests of Special Abilities." *Journal of Experimental Psychology*, 10, (1927), pp. 69-113.
65. Lincoln, Edward A. "The Unreliability of Reliability Coefficients." *Journal of Educational Psychology*, 23, (1932), pp. 11-14.
66. Lincoln, Edward A. "Reliability Coefficients are Still Unreliable." *Journal of Educational Psychology*, 24, (1933), pp. 235-236.
67. Lindquist, E. F., and Cook, Walter W. "Experimental Procedures in Test Evaluation." *Journal of Experimental Education*, 1, (1933), pp. 163-185.
68. Lowell, Frances E. "A Study of the Variability of I.Q.'s in Retests." *Journal of Applied Psychology*, 25, (1941), pp. 341-356.
69. Mangold, Sister Mary Cecilia. *Methods for Measuring the Reliability of Tests.* Catholic University of America, Educational Research Bulletins, Vol. 2. Washington: Catholic Education Press. 1927. Pp. 32.
70. McCall, William A. *Measurement.* New York: Macmillan and Company. 1939. Pp. xv+535.
71. Monroe, Walter Scott. *An Introduction to the Theory of Educational Measurements.* Boston: Houghton Mifflin Co. 1923. Pp. xxiii+364.
72. Mosier, Charles I. "Psychophysics and Mental Test Theory; Fundamental Postulates and Elementary Theorems." *Psychological Review*, 47, (1940), pp. 355-366.
73. Muenzinger, Karl F. "Critical Note on the Reliability of a Test." *Journal of Educational Psychology*, 18, (1927), pp. 424-428.
74. Neyman, J., and Pearson, E. S. "On the Problem of the Most Efficient Test of Statistical Hypotheses." *Philosophical Transactions.* London: Royal Society. Vol. 231-A, (1933), pp. 289-337.
75. Neyman, J. "Sur la vérification des hypothèses statistiques composées." *Bul. Soc. Math. Fr.*, 53, (1935), pp. 1-20.
76. Neyman, J., and Pearson, E. S. "Sufficient Statistics and Uniformly Most Powerful Tests of Statistical Hypotheses." *Statistical Research Memoirs.* London: Department of Statistics, University College. 1936.
77. Neyman, J. "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." *Philosophical Transactions.* London: Royal Society. Vol. 236-A, (1937), pp. 333-380.
78. Otis, Arthur S. *Statistical Method in Educational Measurement.* New York: World Book Co. 1925. Pp. xi+339.
79. Paulsen, G. B. "A Coefficient of Trait Variability." *Psychological Bulletin*, 28, (1931), pp. 218-219.
80. Pearson, K. *Tables of Incomplete Beta-Function.* London: Biometrika Office. 1934.

81. Read, Cecil B. "A Note on Reliability by the Chance Halves Method." *Journal of Educational Psychology*, 30, (1939), pp. 703-704.

82. Remmers, H. H., Shock, N. W., and Kelley, T. L. "An Empirical Study of the Validity of the Spearman-Brown Formula as Applied to the Purdue Rating Scale." *Journal of Experimental Psychology*, 18, (1927), pp. 187-195.

83. Remmers, H. H. "The Equivalence of Judgements to Test Items in the Sense of the Spearman-Brown Formula." *Journal of Educational Psychology*, 22, (1931), pp. 66-71.

84. Remmers, H. H., and Whistler, Laurence. "Test Reliability as a Function of the Method of Computation." *Journal of Educational Psychology*, 29, (1938), pp. 81-92.

85. Remmers, H. H., Karslake, Ruth, and Gage, N. L. "Reliability of Multiple Choice Measuring Instruments as a Function of the Spearman-Brown Prophecy Formula." *Journal of Educational Psychology*, 31, (1940), pp. 583-590.

86. Remmers, H. H., and Ewart, Edwin. "Reliability of Multiple-Choice Measuring Instruments as a Function of the Spearman-Brown Prophecy Formula III." *Journal of Educational Psychology*, 32, (1941), pp. 67-72.

87. Richardson, M. W. "Notes on the Rationale of Item Analysis." *Psychometrika*, 1, No. 1, (1936), pp. 69-76.

88. Richardson, M. W., and Kuder, G. F. "The Calculation of Test Reliability Coefficients Based on the Method of Rational Equivalence." *Journal of Educational Psychology*, 30, (1939), pp. 681-687.

89. Ruch, G. M., and Stoddard, George D. "Comparative Reliabilities of Five Types of Objective Examinations." *Journal of Educational Psychology*, 16, (1925), pp. 89-103.

90. Ruch, G. M., Ackerson, Lutton, and Jackson, Jesse D. "An Empirical Study of the Spearman-Brown Formula as Applied to Educational Test Material." *Journal of Educational Psychology*, 17, (1926), pp. 309-313.

91. Rulon, Phillip J." A Graph for Estimating Reliability in One Range, Knowing it in Another." *Journal of Educational Psychology*, 21, (1930), pp. 140-142.

92. Rulon, Phillip J. "A Simplified Procedure for Determining the Reliability of a Test by Split-halves." *Harvard Educational Review*, 9, (1939), pp. 99-103.

93. Sandiford, Peter. *Educational Psychology*. London: Longmans, Green & Co. 1937. Pp. xix+406.

94. Shen, E. "The Standard Error of Certain Estimated Coefficients of Correlation." *Journal of Educational Psychology*, 15, (1924), pp. 462-465.

95. Shen, E. "A Note on the Standard Error of the Spearman-Brown Formula." *Journal of Educational Psychology*, 17, (1926), pp. 93-94.

96. Sims, Verner Martin. "The Reliability and Validity of Four Types of Vocabulary Tests." *Journal of Educational Research*, 20, (1929), pp. 91-96.

97. Sims, Verner Martin, and Knox, L. B. "The Reliability and Validity of Multiple Response Tests when Presented Orally." *Journal of Educational Psychology*, 23, (1932), pp. 656-662.

98. Skaggs, E. B. "Some Critical Comments on Certain Prevailing Concepts and Methods Used in Mental Testing." *Journal of Applied Psychology*, 11, (1927), pp. 503-508.

99. Slocombe, Charles S. "The Spearman Prophecy Formula." *Journal of Educational Psychology*, 18, (1927), pp. 125-126.
100. Slocombe, Charles S. "A Further Note on the Spearman Prophecy Formula; A Correction." *Journal of Educational Psychology*, 18, (1927), pp. 347-348.
101. Slocombe, Charles S. "The Constancy of 'g', General Intelligence." *British Journal of Psychology*, 17, (1926), pp. 93-110.
102. Symonds, P. M. "A Study of Extreme Cases of Unreliability." *Journal of Educational Psychology*, 15, (1924), pp. 99-106.
103. Symonds, P. M. "Factors Influencing Test Reliability." *Journal of Educacational Psychology*, 19, (1928), pp. 73-87.
104. Symonds, P. M. "Choice of Items for a Test on the Basis of Difficulty." *Journal of Educational Psychology*, 20, (1929), pp. 481-493.
105. Spearman, C. "The Proof and Measurement of Association Between Two Things." *American Journal of Psychology*, 15, (1904), pp. 72-101.
106. Spearman, C. "Correlation Calculated with Faulty Data." *British Journal of Psychology*, 3, (1910), pp. 271-295.
107. Spearman, C. "Correlations of Sums and Differences." *British Journal of Psychology*, 5, (1913), p. 419.
108. Spearman, C. *The Abilities of Man: Their Nature and Measurement.* London: Macmillan and Company. 1932. Pp. x+415+xxxiii.
109. Stephenson, W. "Factorizing the Reliability Coefficient." *British Journal of Psychology*, 25, (1934), pp. 211-216.
110. Stouffer, Samuel A. "Reliability Coefficients in a Correlation Matrix." *Psychometrika*, 1, No. 2, (1936), pp. 17-20.
111. Thomson, Godfrey H. "Weighting for Battery Reliability and Prediction." *British Journal of Psychology*, 29, (1939), pp. 288-306.
112. Thomson, Godfrey H. *The Factorial Analysis of Human Ability.* London: University of London Press. 1939. Pp. xv+326.
113. Thouless, R. H. "The Effect of Errors of Measurement on Correlation Coefficients." *British Journal of Psychology*, 29, (1939), pp. 383-403.
114. Thouless, R. H. "Test Unreliability and Function Fluctuation." *British Journal of Psychology*, 26, (1936), p. 325.
115. Thurstone, L. L. "A Note on the Spearman-Brown Formula." *Journal of Experimental Psychology*, 11, (1928), pp. 62-63.
116. Thurstone, L. L. *The Reliability and Validity of Tests.* Ann Arbor: Edwards Bros. 1933. Pp. 113.
117. Thurstone, Thelma G. "The Difficulty of a Test and its Diagnostic Value." *Journal of Educational Psychology*, 23, (1932), pp. 335-343.
118. Trimble, Otis C. "The Oral Examination: Its Validity and Reliability." *School and Society*, 39, (1934), pp. 550-552.
119. Turney, Austin H. "The Cumulative Reliability of Frequent Short Objective Tests." *Journal of Educational Research*, 25, (1932), pp. 290-295.
120. Walker, Helen M. *Studies in the History of Statistical Method, With Special Reference to Certain Educational Problems.* Baltimore: The Williams and Wilkins Company, 1929. Pp. viii+229.
121. Weidmann, C. C. "Reliability or Consistency Coefficient." *School and Society*, 31, (1930), p. 674.

122. Weidmann, Charles C., and Newens, Lyndall Fisher. "The Effect of Directions Preceding True-False and Indeterminate Statement Examinations upon Distributions of Test Scores." *Journal of Educational Psychology*, 24, (1933), pp. 97-106.
123. Woodrow, H. "Quotidian Variability." *Psychological Review*, 39, (1932), pp. 245-256.
124. Wherry, Robert J. "The Shrinkage of the Brown-Spearman Prophecy Formula." *Annals of Mathematical Statistics*, 6, (1935), pp. 183-189.

THE DEPARTMENT OF EDUCATIONAL RESEARCH

UNIVERSITY OF TORONTO
371 BLOOR STREET WEST, TORONTO 5
CANADA

BULLETINS

No.

1—On the Counting of New Words in Textbooks for Teaching Foreign Languages.
E. SWENSON & M. P. WEST. 1934. Pp. 40 Price **50c.**

2—A Critical Examination of Basic English.
M. P. WEST, E. SWENSON et al. Pp. 53. 1934. (Out of Print).

3—The V
JOH

4—Defi
MIC

5—The **nce Using**
C.

6—Cor
JOH

7—The
MA

8—For
P.
Pric **. Pp. 93.**

9—The
C.

10—Bib
A.

11—Ap **d to Edu-**
Ro

12—Stu
Ro **00.**

42